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

Assignment 4 - Due date 02/11/25

Mazhar Bhuyan

## 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.\

```
#Load/install required package here

library(dplyr)
library(forecast)
library(ggplot2)
library(tidyverse)
library(cowplot)
library(stargazer)
library(readxl)
library(trend)
library(tseries)
library(Kendall)
```

## 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
Energy_Data <- read_excel("./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx
  skip = 12,
  sheet = "Monthly Data",
  col_names = FALSE)
#head(Energy_Data)

col_names <- read_excel("./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
      skip=10,
      n_max = 1,
      sheet="Monthly Data",
      col_names = FALSE)
#head(col_names)

colnames(Energy_Data) <- col_names

data <- Energy_Data[,c(1,5)]
nobs <- nrow(Energy_Data)
#Create vector t - time index
t <- 1:nobs

Renewable_Energy_Data_ts <- ts(data[t,2], frequency=12,start=c(1973,1))
#head(Renewable_Energy_Data_ts)
```

## 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.

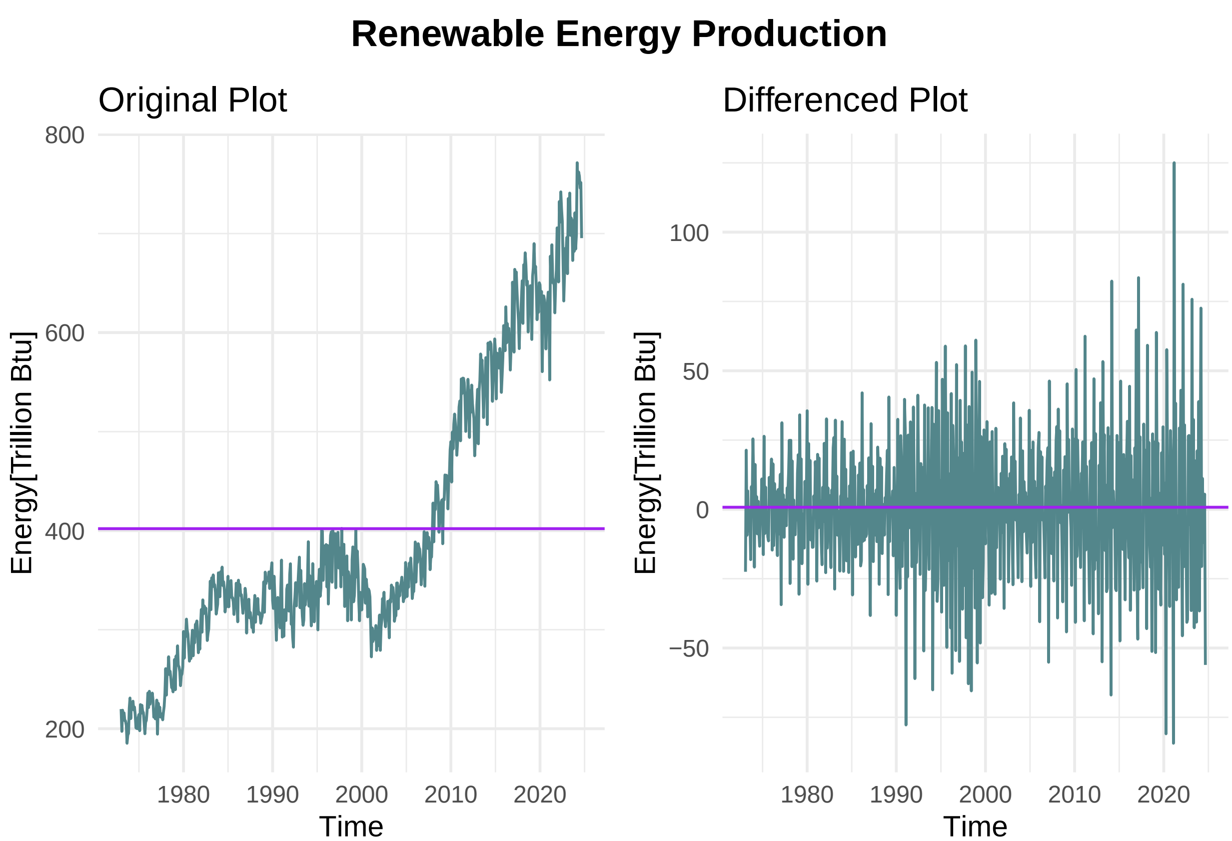
Try differencing at lag 1 only once, i.e., make `lag=1` and `differences=1`. Plot the differenced series. Do the series still seem to have trend?

```r
Renewable_Production_diff <- diff(Renewable_Energy_Data_ts,
                                  lag = 1,
                                  differences = 1
)

ts_plot <- autoplot(Renewable_Energy_Data_ts,
                    color = "cadetblue4") +
          xlab("Time") +
          ylab("Energy[Trillion Btu]") +
          geom_hline(yintercept = mean(Renewable_Energy_Data_ts),
                     color = "purple",
                     linetype = "solid") +
          ggtitle("Original Plot")+
          theme_minimal()
#print(ts_plot)
```

```r
diff_plot <- autoplot(Renewable_Production_diff,
                      color = "cadetblue4") +
          xlab("Time") +
          ylab("Energy[Trillion Btu]")+
          geom_hline(yintercept = mean(Renewable_Production_diff),
                          color = "purple",
                          linetype = "solid") +
          ggtitle("Differenced Plot") +
          theme_minimal()
#print(diff_plot)

plot_comparison <- plot_grid(ts_plot,diff_plot,nrow=1,ncol=2)
#print(plot_comparison)
title <- ggdraw() +
          draw_label("Renewable Energy Production", fontface='bold')
plot_grid(title,plot_comparison,nrow=2,ncol=1,rel_heights = c(0.1,1))
```




Answer: After differencing at lag 1, difference 1, it seems like that the strong upward trend is removed. The values fluctuate around zero indicating a more stationary behavior. But the plot also shows very long spikes at the latest times. It might require more investigation to find out what is going on.

### 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.

```
## Linear Fit Model
lm_renewable=lm(Renewable_Energy_Data_ts~t)
beta0_renewable=lm_renewable$coefficients[1]
beta1_renewable=lm_renewable$coefficients[2]
#print(summary(lm_renewable))

stargazer(lm_renewable, type = "text",
          title = "Renewable Energy Production Linear Fit Model",
          align = TRUE, star.cutoffs = c(0.1, 0.05, 0.01))
```

```
##
## Renewable Energy Production Linear Fit Model
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                      Renewable_Energy_Data_ts
## -----------------------------------------------
## t                            0.724***
##                               (0.014)
##
## Constant                    176.873***
##                               (4.962)
##
## -----------------------------------------------
## Observations                   621
## R2                            0.816
## Adjusted R2                   0.816
## Residual Std. Error      61.750 (df = 619)
## F Statistic         2,742.930*** (df = 1; 619)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
## Detrended Series

detrended_renewable <- Renewable_Energy_Data_ts - (beta0_renewable +
                                                    beta1_renewable*t)
detrended_renewable <- ts(detrended_renewable,
                          frequency = 12,
                          start = c(1973,1))

detrend_plot <- autoplot(detrended_renewable) +
                xlab("Time") +
                ylab("Energy[Trillion Btu]")+
                geom_hline(yintercept = mean(detrended_renewable),
                color = "purple",
                linetype = "solid") +
                ggtitle("Detrend Plot") +
```

```r
                theme_minimal()
#print(detrend_plot)
```

### 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.

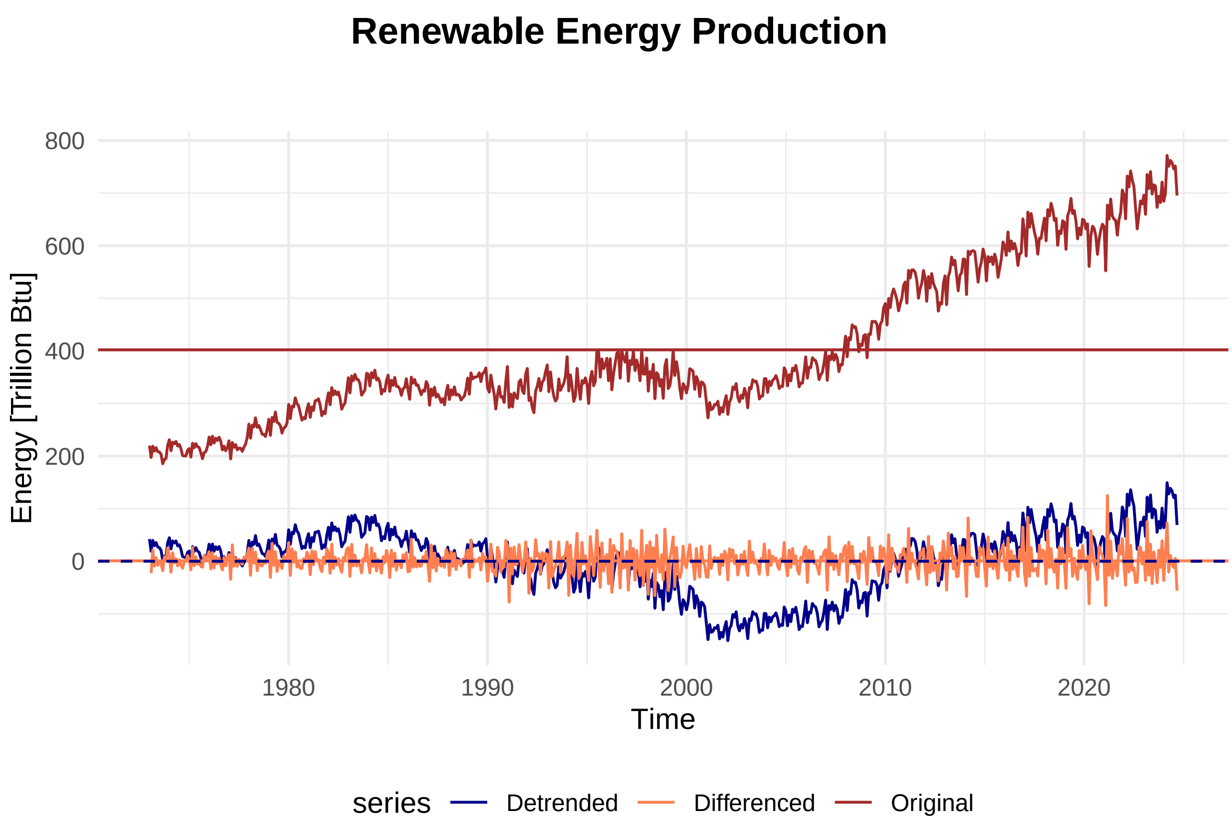
Using autoplot() + autolayer() create a plot that shows the three series together. Make sure your plot has a legend. The easiest way to do it is by adding the `series=` argument to each autoplot and autolayer function. Look at the key for A03 for an example on how to use autoplot() and autolayer().

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

```r
plot_comparison <- autoplot(Renewable_Energy_Data_ts,
                            series = "Original") +
  autolayer(detrended_renewable,
            series = "Detrended") +
  autolayer(Renewable_Production_diff,
            series = "Differenced") +
  scale_color_manual(
    values = c("Original" = "brown",
               "Detrended" = "darkblue",
               "Differenced" = "coral")) +
  ggtitle("") +
  xlab("Time") +
  ylab("Energy [Trillion Btu]") +
  geom_hline(yintercept = mean(Renewable_Energy_Data_ts), color = "brown", linetype = "solid") +
  geom_hline(yintercept = mean(Renewable_Production_diff), color = "coral", linetype = "solid") +
  geom_hline(yintercept = mean(detrended_renewable), color = "darkblue", linetype = "dashed") +
  theme_minimal() +
  theme(legend.position = "bottom")

#print(plot_comparison)

title <- ggdraw() +
          draw_label("Renewable Energy Production", fontface='bold')
plot_grid(title,plot_comparison,nrow=2,ncol=1,rel_heights = c(0.1,1))
```




Answer: The Original Series showed a strong upward trend, which was partially removed by detrending. However, the detrended series still exhibits significant variation and does not fluctuate around zero. This indicates that the detrending process only partially removed the trend. From the detrended plot, we can observe that the series began trending upward again around the year 2000.

In contrast, first-order differencing successfully removed the trend, as the series now fluctuates 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?

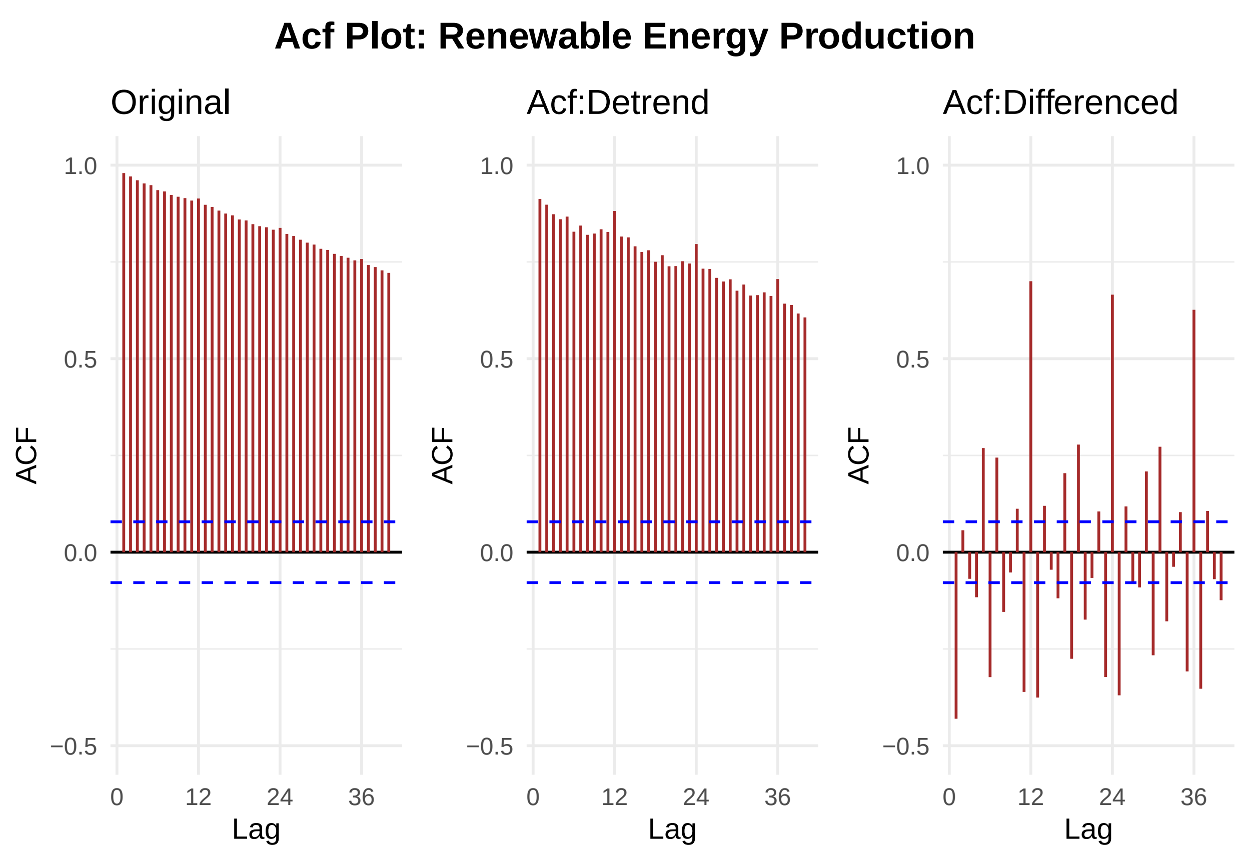
```
acf_renewable_ts <- ggAcf(Renewable_Energy_Data_ts, lag.max = 40,
                          color = "brown")+
                    ggtitle("Original") +
                    ylim(c(-0.5,1)) +
                    theme_minimal()
#print(acf_renewable_ts)

acf_detrended_renewable <- ggAcf(detrended_renewable,lag.max = 40,
```

```
                                 color = "brown") +
                            ggtitle("Acf:Detrend") +
                            ylim(c(-0.5,1)) +
                            theme_minimal()
#print(acf_detrended_renewable)

acf_diff_renewable <-ggAcf(Renewable_Production_diff, lag.max = 40,
                           color = "brown") +
                      ggtitle("Acf:Differenced") +
                      ylim(c(-0.5,1)) +
                      theme_minimal()
#print(acf_diff_renewable)

acf_plot_comparison <- plot_grid(acf_renewable_ts,acf_detrended_renewable,
                                 acf_diff_renewable,
                                nrow=1,ncol=3)
#print(acf_plot_comparison)
title <- ggdraw() +
          draw_label("Acf Plot: Renewable Energy Production", fontface='bold')
plot_grid(title,acf_plot_comparison,nrow=2,ncol=1,rel_heights = c(0.1,1))
```




Answer: The original series shows strong autocorrelation, as evidenced by the slow decay over time. This indicates that past values significantly impact present values. After fitting a linear trend model, we observe a spike every 12 lags, suggesting that the series follows a pattern that restarts every 12 lags. While the autocorrelation is less pronounced than in the original series, it remains prominent, with every lag being significant.

After applying the first-order difference, the trend component is mostly removed. As observed, most of the lags are not significant, suggesting the presence of white noise. However, the series still contains several significant spikes.

Between the linear trend fit and the first-order difference, the latter effectively removes the trend component.

### 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.

```r
summary(SeasonalMannKendall(Renewable_Energy_Data_ts))
```

```
## Score =  12468 , Var(Score) = 190008
## denominator =  15758.5
## tau = 0.791, 2-sided pvalue =< 2.22e-16
```

```r
print(smk.test(Renewable_Energy_Data_ts))
```

```
##
##  Seasonal Mann-Kendall trend test (Hirsch-Slack test)
##
## data:  Renewable_Energy_Data_ts
## z = 28.601, p-value < 2.2e-16
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##      S   varS
##  12468 190008
```

```r
adf_test <- adf.test(Renewable_Energy_Data_ts)
print(adf_test)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  Renewable_Energy_Data_ts
## Dickey-Fuller = -1.0898, Lag order = 8, p-value = 0.9242
## alternative hypothesis: stationary
```

Answer:

SMK Test: p-value the of the original time series of renewable data is extremely small ($< 0.05$), so we reject the null hypothesis (Ho). This means that there is a statistically significant trend in the time series.

Test Statistic (z = 28.601). A large positive z value suggests an increasing trend in the series.

S Statistic, S = 12468. It suggest an upward trend in the data.

The tau = 0.791 is close to 1, meaning there is a strong correlation between time and increasing values in the series. This confirms that renewable energy production has consistently grown over time.

The Total Renewable Energy Production time series has an increasing trend. This suggests that the production of renewable energy has been steadily rising over time. Since the test confirms a trend, the series is likely non-stationary.

ADF Test: The p-value (0.9242) is very high (> 0.05), meaning we fail to reject the null hypothesis (Ho) meaning we accept the null hypothesis of having an unit root. The time series has a unit root and is not stationary.

> As expected from the plot of Q3 we can see that time series has a stochastic trend which can be removed by differencing.

### 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().

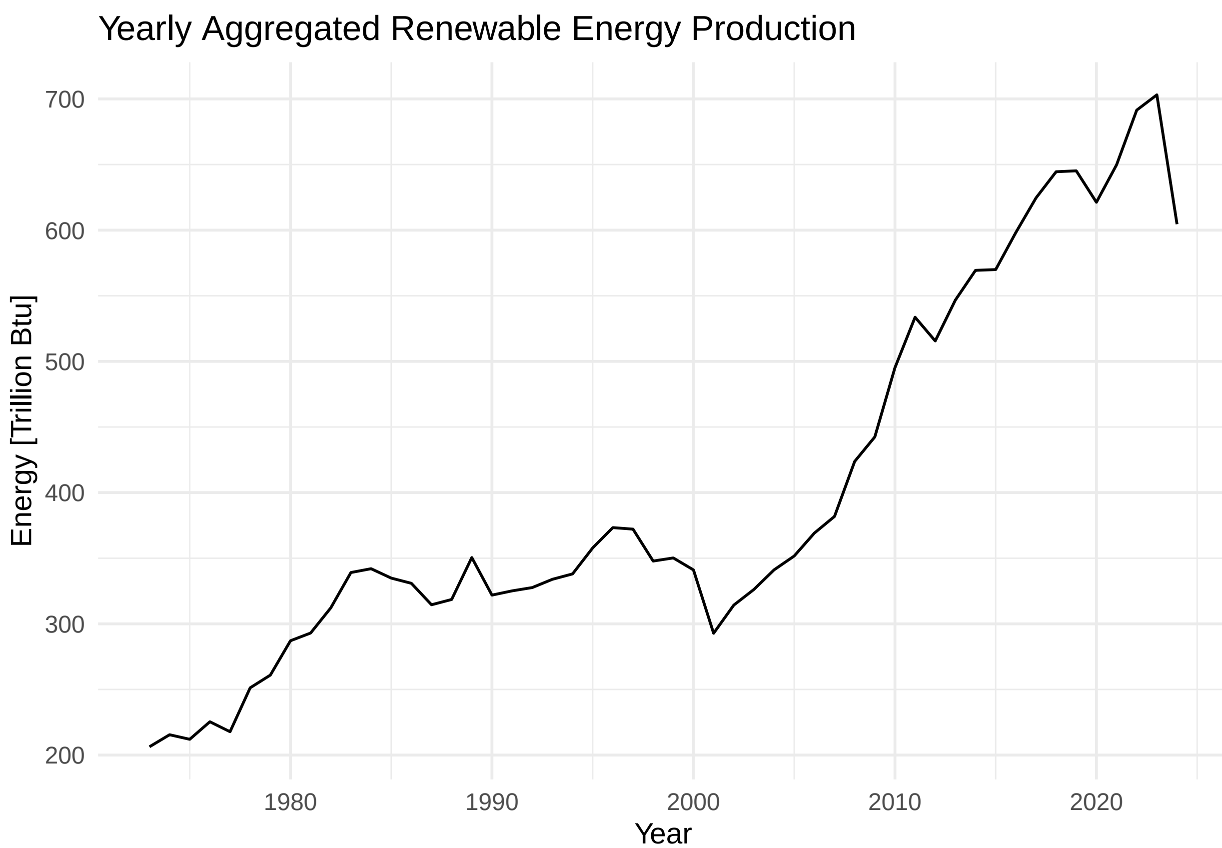
```r
## Converting time series into a matrix (reshape data by year and month)

years <- floor(time(Renewable_Energy_Data_ts))
months <- cycle(Renewable_Energy_Data_ts)

## Creating a matrix format
renewable_matrix <- matrix(Renewable_Energy_Data_ts,
                           nrow = 12,
                           byrow = FALSE)
## Computing yearly means
yearly_means <- colMeans(renewable_matrix, na.rm = TRUE)

## Converting aggregrated yearly data into a time series object

yearly_ts <- ts(yearly_means,
                start = start(Renewable_Energy_Data_ts)[1],
                frequency = 1)
autoplot(yearly_ts) +
  ggtitle("Yearly Aggregated Renewable Energy Production") +
  xlab("Year") +
  ylab("Energy [Trillion Btu]") +
  theme_minimal()
```




**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
# Applying Mann-Kendall Trend Test
summary(MannKendall(yearly_ts))
```

```
## Score =  1070 , Var(Score) = 16059.33
## denominator =  1326
## tau = 0.807, 2-sided pvalue =< 2.22e-16
```

```r
print(mk.test(yearly_ts))
```

```
##
##  Mann-Kendall trend test
##
## data:  yearly_ts
## z = 8.4356, n = 52, p-value < 2.2e-16
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 1.070000e+03 1.605933e+04 8.069382e-01
```

```r
# Applying Spearman Rank Correlation Test
spearman_yearly <- cor.test(time(yearly_ts), yearly_ts, method = "spearman")
print(spearman_yearly)
```

```
##
##  Spearman's rank correlation rho
##
## data:  time(yearly_ts) and yearly_ts
## S = 1908, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.918552
```

```r
# Applying Augmented Dickey-Fuller (ADF) Test
adf_yearly <- adf.test(yearly_ts)
print(adf_yearly)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  yearly_ts
## Dickey-Fuller = -1.6634, Lag order = 3, p-value = 0.7098
## alternative hypothesis: stationary
```

Answer:

The Mann-Kendall test shows a strong increasing trend in the yearly data. The tau value is 0.807, which indicates a significant upward movement. The p-value is less than 2.22e-16, which is much smaller than 0.05. This means we reject the null hypothesis (Ho) and confirm that a trend exists.

The Spearman rank correlation test also supports this finding. The rho value is 0.918552, showing a strong positive correlation over time. The p-value is less than 2.2e-16, which confirms that the trend is statistically significant. This means that as time progresses, renewable energy production has consistently increased.

The Augmented Dickey-Fuller (ADF) test, however, suggests that the data is non-stationary. The p-value is 0.7098, which is much higher than 0.05. Since the p-value is large, we fail to reject the null hypothesis (Ho) of a unit root. This means that the series has a stochastic trend, and differencing is needed to make it stationary.

These results match the findings from the monthly data analysis in Q6. Both the Mann-Kendall and Spearman tests confirm an increasing trend, while the ADF test shows that the data is non-stationary. This means that renewable energy production has been rising over time, but the series needs transformation to remove the trend.