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

Assignment 4 - Due date 02/11/25

Aye Nyein Thu

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(readxl)
library(ggplot2)
library(forecast)
library(Kendall)
library(tseries)
library(lubridate)
library(cowplot)
library(trend)
library(dplyr)

# Set my theme
mytheme <- theme_classic() +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top",
        plot.title = element_text(hjust = 0.5, face = "bold"))

theme_set(mytheme)
```

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"**.

```
# Import Dataset
Energy <- 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
Column_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(Energy) <- Column_Names

# Select the columns of interest and format date
Energy_cleaned <- Energy %>%
  select("Month",
         "Total Renewable Energy Production") %>%
  mutate(Month = as.Date(Month, format = "%Y-%m-%d")) %>%
  rename(Renewable = "Total Renewable Energy Production")

# Assign values to no. of observation
nobs <- nrow(Energy_cleaned)

# Specify starting points
start_year <- as.numeric(format(min(Energy_cleaned$Month), "%Y"))
start_month <- as.numeric(format(min(Energy_cleaned$Month), "%m"))

# Transform data frame to time series objects
Energy_ts <- ts(Energy_cleaned[,2],
                start = c(start_year, start_month), frequency = 12)
```

Stochastic Trend and Stationarity Tests

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

Q1

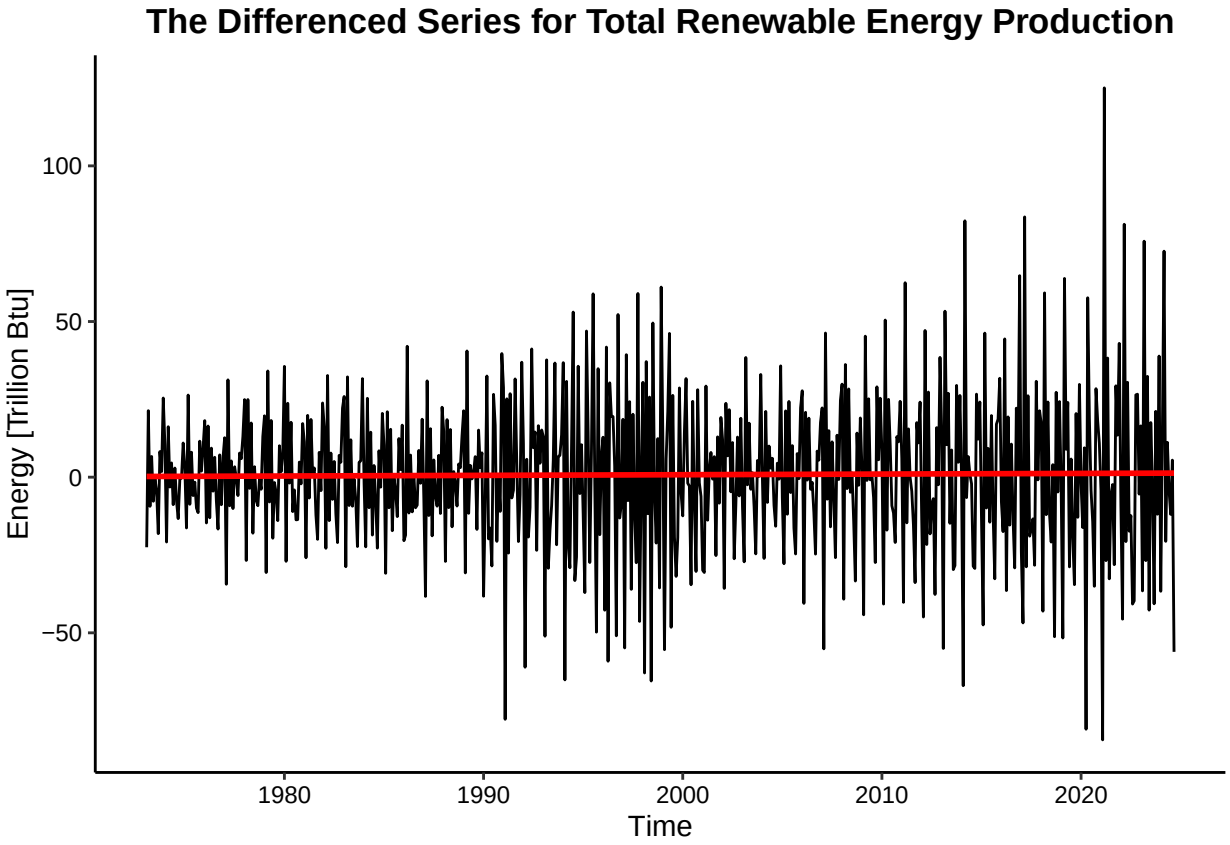
Difference the "Total Renewable Energy Production" series using function diff(). Function diff() is from package base and take three main arguments: * x vector containing values to be differenced; * *lag* integer indicating with lag to use; * *differences* integer indicating how many times series should 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
# Difference the series
diff_renew <- diff(Energy_ts, lag = 1, differences = 1)
class(diff_renew)
```

```
## [1] "ts"
```

```r
# Plot the differenced series
autoplot(diff_renew) +
  ylab("Energy [Trillion Btu]") +
  ggtitle("The Differenced Series for Total Renewable Energy Production") +
  geom_smooth(method = "lm", se = FALSE, color = "red")
```



Answer: When the total renewable energy production is transformed into the differenced series, the series do not exhibit a significant upward or downward trend. The fluctations are more in random nature, rather than representing trends.

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.

```r
# Create vector t
t <- c(1:nobs)

# Fit a linear model for Renewable
linear_renew <- lm(Energy_cleaned$Renewable ~ t)
summary(linear_renew)
```

```
##
## Call:
## lm(formula = Energy_cleaned$Renewable ~ t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -151.11  -37.84   13.53   41.76  149.42
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 176.87293    4.96189   35.65   <2e-16 ***
## t             0.72393    0.01382   52.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61.75 on 619 degrees of freedom
## Multiple R-squared:  0.8159, Adjusted R-squared:  0.8156
## F-statistic:  2743 on 1 and 619 DF,  p-value: < 2.2e-16
```

```r
# Save the regression coefficients for Renewable
beta0_linear_renew <- as.numeric(linear_renew$coefficients[1])
beta1_linear_renew <- as.numeric(linear_renew$coefficients[2])

# Create detrended series from the linear model for renewable
linear_trend_renew <- beta0_linear_renew + beta1_linear_renew * t
ts_linear_trend_renew <- ts(linear_trend_renew,
                            start=c(start_year,start_month), frequency=12)

detrend_renew <- Energy_cleaned[,2] - linear_trend_renew
ts_detrend_renew <- ts(detrend_renew,
                       start=c(start_year,start_month), frequency = 12)
```

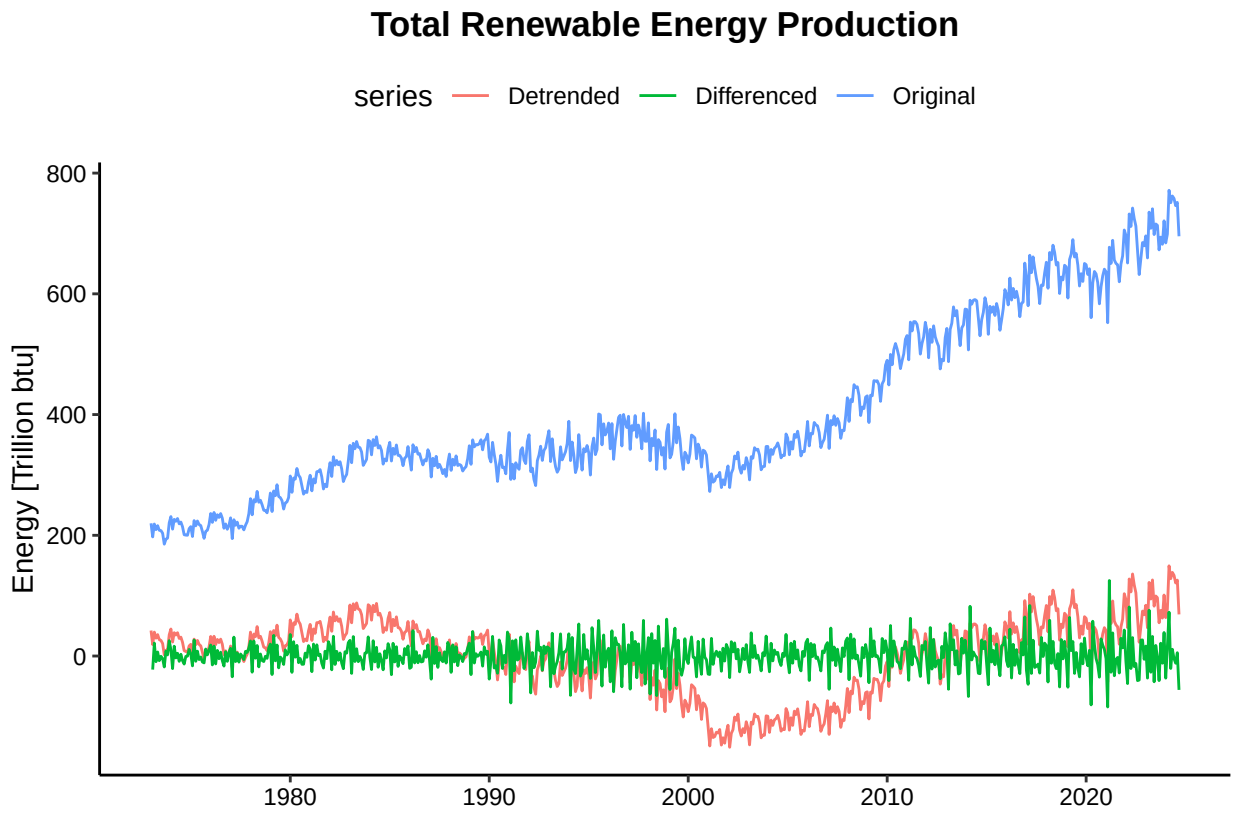
Q3

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 original, detrended and differenced series
autoplot(Energy_ts, series = "Original") +
  autolayer(ts_detrend_renew, series = "Detrended") +
  autolayer(diff_renew, series = "Differenced") +
  labs(title = "Total Renewable Energy Production",
       x = "", y = "Energy [Trillion btu]")
```



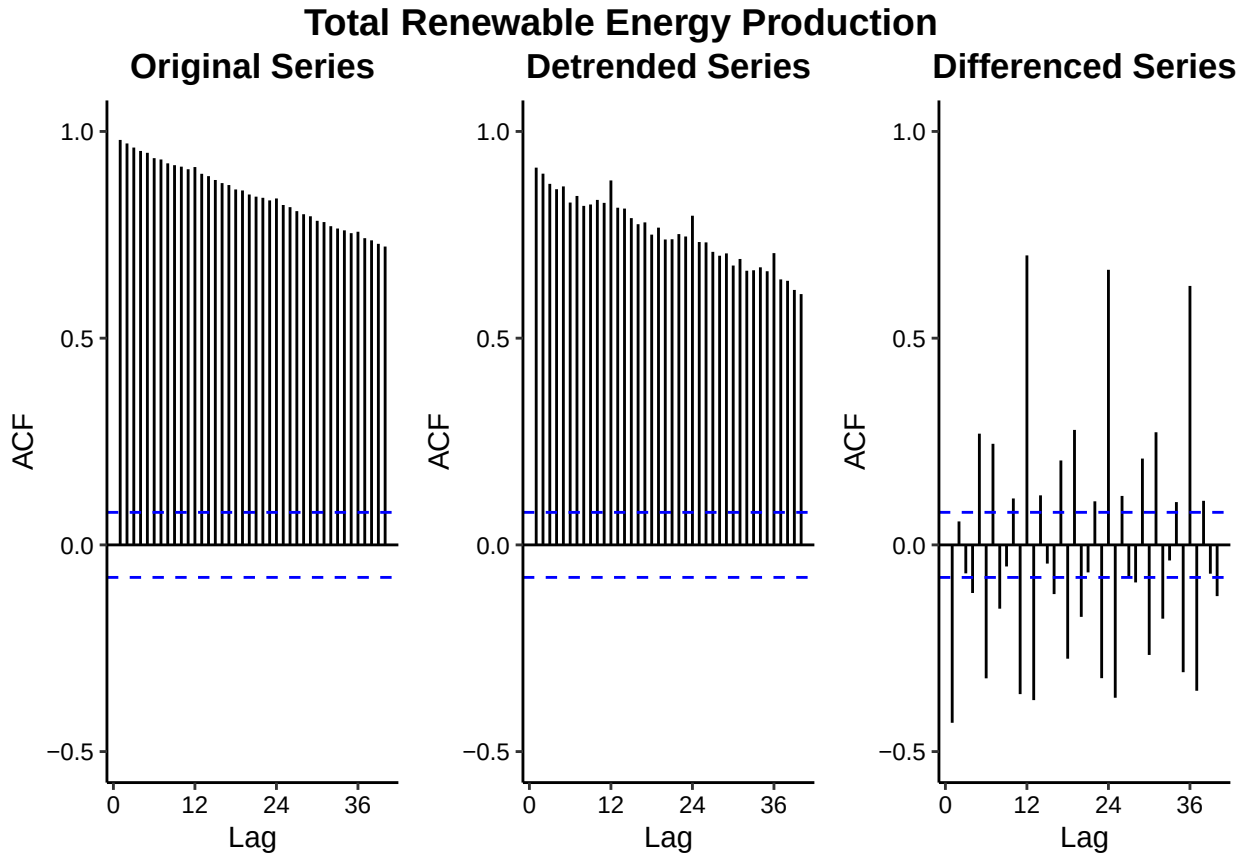
Answer: When the original, detrended and differences series are plotted together, it is apparent that the differenced series is more efficient in removing the trend. The original series has a significant upward trend. The detrended series with the use of linear regression model still exhibits a slight trend component. When the differencing method is used, the series do not showcase an apparent upward or downward trend anymore. Therefore, we could assume that the differenced series is more efficient in removing the trend for total renewable energy production data.

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?

```r
# Plot ACF for three series
plot_grid(
```

```
  autoplot(Acf(Energy_ts, lag.max = 40, plot = FALSE),
           main = "Original Series", ylim = c(-0.5,1)),
  autoplot(Acf(ts_detrend_renew, lag.max = 40, plot = FALSE),
           main = "Detrended Series", ylim = c(-0.5,1)),
  autoplot(Acf(diff_renew, lag.max = 40, plot = FALSE),
           main = "Differenced Series", ylim = c(-0.5,1)),
  nrow = 1, ncol = 3
) +
  ggtitle("Total Renewable Energy Production") +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



Answer: As per the ACF plots, both original series and linear regression represent a strong trend and strong autocorrelation on the past values. These patterns are no longer present in the differenced series as it effectively take out the trend component. Therefore, the differencing is more efficient than linear model.

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
# Run Seasonal Mann-Kendall test
summary(SeasonalMannKendall(Energy_ts))
```

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

```r
summary(smk.test(Energy_ts))
```

```
##
##  Seasonal Mann-Kendall trend test (Hirsch-Slack test)
##
## data: Energy_ts
## alternative hypothesis: two.sided
##
## Statistics for individual seasons
##
## H0
##                        S    varS   tau     z   Pr(>|z|)
## Season 1:   S = 0  1036 16059.3 0.781 8.167 3.1546e-16 ***
## Season 2:   S = 0  1038 16059.3 0.783 8.183 2.7676e-16 ***
## Season 3:   S = 0  1024 16059.3 0.772 8.073 6.8833e-16 ***
## Season 4:   S = 0  1026 16059.3 0.774 8.088 6.0477e-16 ***
## Season 5:   S = 0  1026 16059.3 0.774 8.088 6.0477e-16 ***
## Season 6:   S = 0  1042 16059.3 0.786 8.215 < 2.22e-16 ***
## Season 7:   S = 0  1072 16059.3 0.808 8.451 < 2.22e-16 ***
## Season 8:   S = 0  1090 16059.3 0.822 8.593 < 2.22e-16 ***
## Season 9:   S = 0  1066 16059.3 0.804 8.404 < 2.22e-16 ***
## Season 10:   S = 0 1023 15158.3 0.802 8.301 < 2.22e-16 ***
## Season 11:   S = 0 1016 15157.3 0.797 8.244 < 2.22e-16 ***
## Season 12:   S = 0 1009 15158.3 0.791 8.187 2.6740e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Run ADF test
print(adf.test(Energy_ts, alternative = "stationary"))
```

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

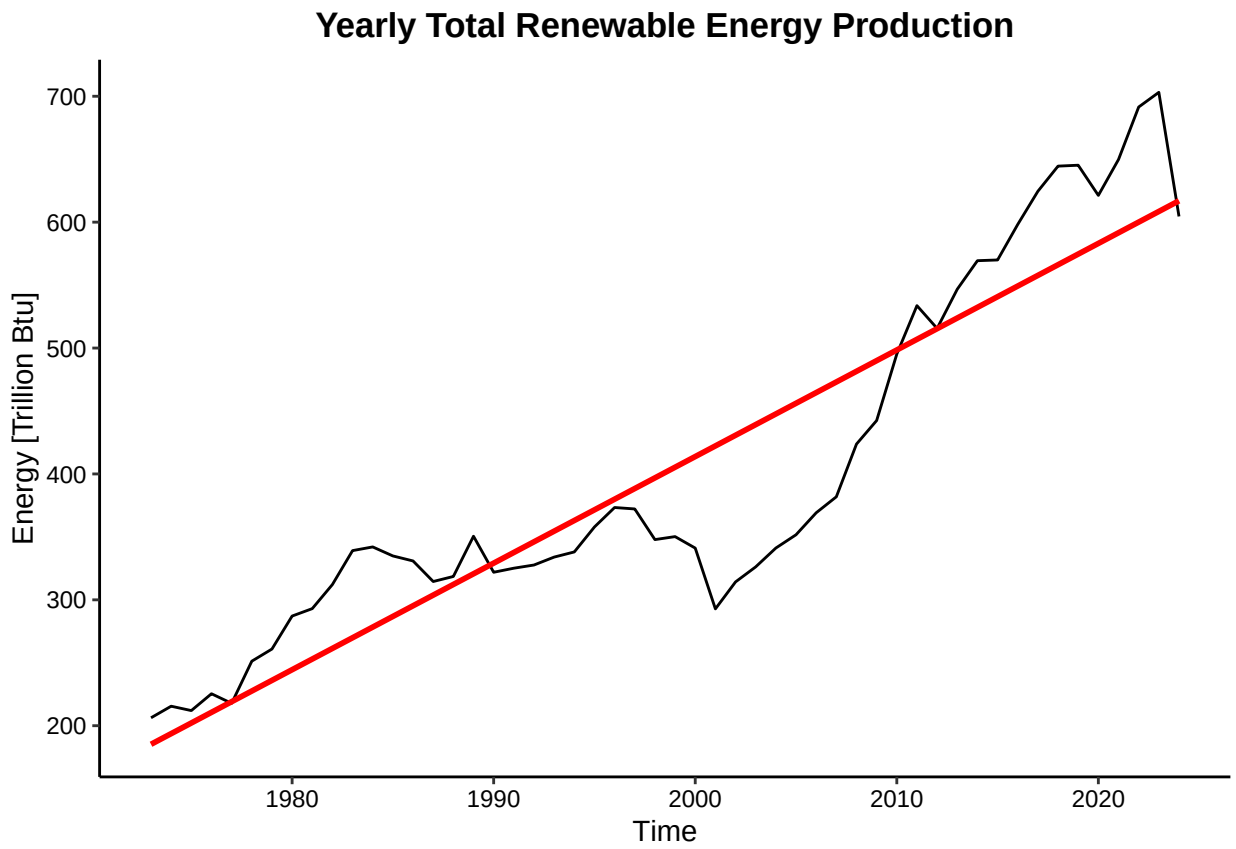
Answer: The seasonal Mann-Kendall tests indicate that the original series of total renewable energy production has a very strong increasing trend at tau value (0.79) and score (12468). Since the p-value is less than 0.05, it is statistically significant that the data series has a trend.

The Augmented Dickey-Fuller (ADF) test shows that the series has a ADF value at -1.09 and p-value at 0.92. Since the value is approximately equal to 1 and p-value is higher than 0.05, we failed to reject the null hypothesis. Therefore, we do not have enough evidence to reject the hypothesis that the series has a unit root or stochastic trend. As we could not rule out the existence of stochastic trend, it is justifiable to use teh differencing method to remove the 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().

```
# Store series in a matrix
Energy_matrix <- matrix(Energy_ts, byrow = FALSE, nrow = 12)

# Take the columns mean
Energy_yearly <- colMeans(Energy_matrix)

# Create the year vector
my_year <- c(1973:2024)

# Create data frame
Energy_yearly_data <- data.frame(my_year,"Renew_yearly"=Energy_yearly)

# Convert the data frame to time series
ts_Energy_yearly <- ts(Energy_yearly_data[,2],
                       start = c(start_year,start_month), frequency = 1)

# Plot the series
autoplot(ts_Energy_yearly) +
  ylab("Energy [Trillion Btu]") +
  ggtitle("Yearly Total Renewable Energy Production") +
  geom_smooth(method = "lm", se = FALSE, color = "red")
```



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?

```
# Run Mann Kendall Test
summary(MannKendall(ts_Energy_yearly))
```

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

```
# Run Spearman Correlation Test
print(cor.test(Energy_yearly_data$Renew_yearly, my_year, method="spearman"))
```

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

```r
# Run ADF Test
print(adf.test(ts_Energy_yearly,alternative = "stationary"))
```

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

Answer: The Mann Kendall test based on yearly mean data provides enough evidence that the time series has an increasing trend at tau value (0.81) and it is statistically signficiant at 5% level. The Spearman's correlation results also indicates the existence of increasing monotonic trend at rho (0.92) and p-value less than 0.05.

Finally, the ADF test with the use of yearly mean also generates the same conclusion as Question 6. Since the p-value is greather than 0.05, we could reject the null hypothesis that the yearly time series has a unit root. Therefore, the value (-1.66) is slightly far away from 1 compared to Question 6, the p-value at 0.71 exhibits that the series has the probability of the stochastic trend.

To conclude, even if we remove the seasonality and aggregate data on a yearly basis, we still have a strong and significant increasing trend and could not rule out the possibility of stochastic trend. Therefore, the differencing method is more efficient than the linear model to remove the trend component in the total renewable energy production series.