ENV 797 - Time Series Analysis for Energy and Environment Applications | Spring 2026

Assignment 4 - Due date 02/10/26

Yeeun Kim

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_Sp26.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(forecast)
library(tseries)
library(Kendall)
library(dplyr)
library(readxl)
library(openxlsx)
library(ggplot2)
library(Kendall)
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 December 2025 Monthly Energy Review. **For this assignment you will work only with the column "Total Renewable Energy Production"**.

```
#Importing data set - you may copy your code from A3
#Importing data
energy_data1 <- read_excel(path="/Users/yeeunkim/Library/CloudStorage/OneDrive-DukeUniversity/2026 Spri
```

```r
# Extract the column names from row 11
read_col_names <- read_excel(path="/Users/yeeunkim/Library/CloudStorage/OneDrive-DukeUniversity/2026 Sp

#Assign the column names to the data set
colnames(energy_data1) <- read_col_names

#Visualize the first rows of the data set
head(energy_data1)
```

```
## # A tibble: 6 x 14
##   Month               `Wood Energy Production` `Biofuels Production`
##   <dttm>                                 <dbl> <chr>
## 1 1973-01-01 00:00:00                     130. Not Available
## 2 1973-02-01 00:00:00                     117. Not Available
## 3 1973-03-01 00:00:00                     130. Not Available
## 4 1973-04-01 00:00:00                     125. Not Available
## 5 1973-05-01 00:00:00                     130. Not Available
## 6 1973-06-01 00:00:00                     125. Not Available
## # i 11 more variables: `Total Biomass Energy Production` <dbl>,
## #   `Total Renewable Energy Production` <dbl>,
## #   `Hydroelectric Power Consumption` <dbl>,
## #   `Geothermal Energy Consumption` <dbl>, `Solar Energy Consumption` <chr>,
## #   `Wind Energy Consumption` <chr>, `Wood Energy Consumption` <dbl>,
## #   `Waste Energy Consumption` <dbl>, `Biofuels Consumption` <chr>,
## #   `Total Biomass Energy Consumption` <dbl>, ...
```

```r
#Select the columns
energy_select <- energy_data1 %>%
  select(1,5)
head(energy_select)
```

```
## # A tibble: 6 x 2
##   Month               `Total Renewable Energy Production`
##   <dttm>                                            <dbl>
## 1 1973-01-01 00:00:00                                220.
## 2 1973-02-01 00:00:00                                197.
## 3 1973-03-01 00:00:00                                219.
## 4 1973-04-01 00:00:00                                209.
## 5 1973-05-01 00:00:00                                216.
## 6 1973-06-01 00:00:00                                208.
```

```r
#TS data
ts_energy_data <- ts(energy_select[,2],start = c(1973,1), frequency = 12)
```

Stochastic Trend and Stationarity Tests

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

Q1

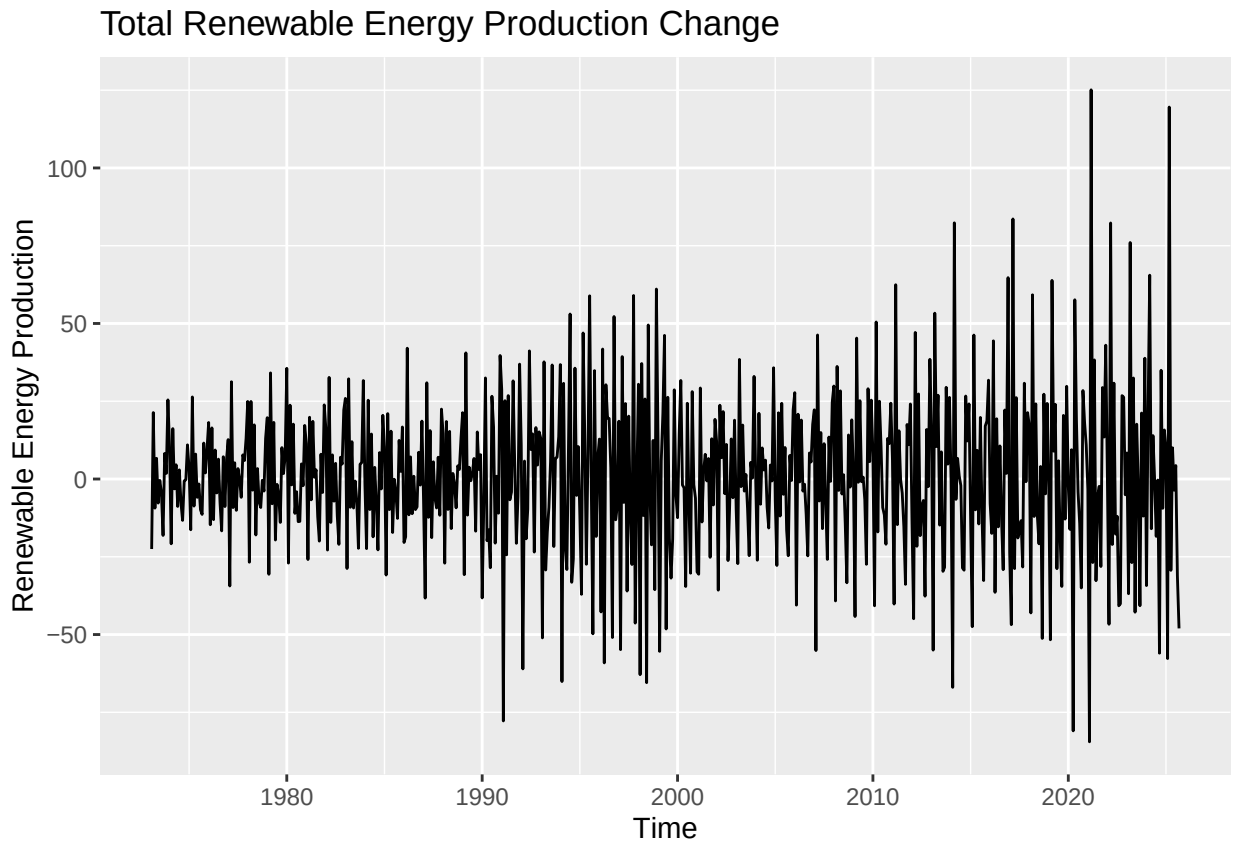
Difference the "Total Renewable Energy Production" series using function diff(). Function diff() is from package base and take three main arguments: * *x* vector containing values to be differenced; * *lag* integer indicating with lag to use; * *differences* integer indicating how many times series should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

> Answer: Although the differenced series fluctuate upward at certain periods, it does not show an increasing trend, as it oscillates around a constant mean around zero.

```
# Differencing
diff_renew <- diff(ts_energy_data, lag=1, differences=1)
ts_diff_renew <- ts(diff_renew, start = c(1973,2), frequency = 12) #It starts from february due to the

# Making a plot
autoplot(diff_renew) +
  labs(
    title = "Total Renewable Energy Production Change",
    x = "Time",
    y = "Renewable Energy Production",
    color = NULL)
```



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 assign same name for the time series object that you had in A3, otherwise the code will not work.

```
#Create vector t
t <- 1:length(ts_energy_data[,1])

#Renewable_Linear model
renew_linear_model <- lm(ts_energy_data[,1]~t)
summary(renew_linear_model)
```

```
##
## Call:
## lm(formula = ts_energy_data[, 1] ~ t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -154.81  -39.55   12.52   41.49  171.15
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 171.44868    5.11085   33.55   <2e-16 ***
## t             0.74999    0.01397   53.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.22 on 631 degrees of freedom
## Multiple R-squared:  0.8204, Adjusted R-squared:  0.8201
## F-statistic:  2883 on 1 and 631 DF,  p-value: < 2.2e-16
```

```
#Renewable_Store regression coefficients
beta0=as.numeric(renew_linear_model$coefficients[1])  #first coefficient is the intercept term or beta0
beta1=as.numeric(renew_linear_model$coefficients[2])  #second coefficient is the slope or beta1

#Renewable: remove the trend from series
renew_detrend <- energy_select[,2]-(beta0+beta1*t)

#Renewable: note detrend will be a data frame and not a ts object
class(renew_detrend)
```

```
## [1] "data.frame"
```

```
#Renewable: Transfer into ts object
ts_renew_detrend <- ts(renew_detrend,frequency=12,start=c(1973,1))
```

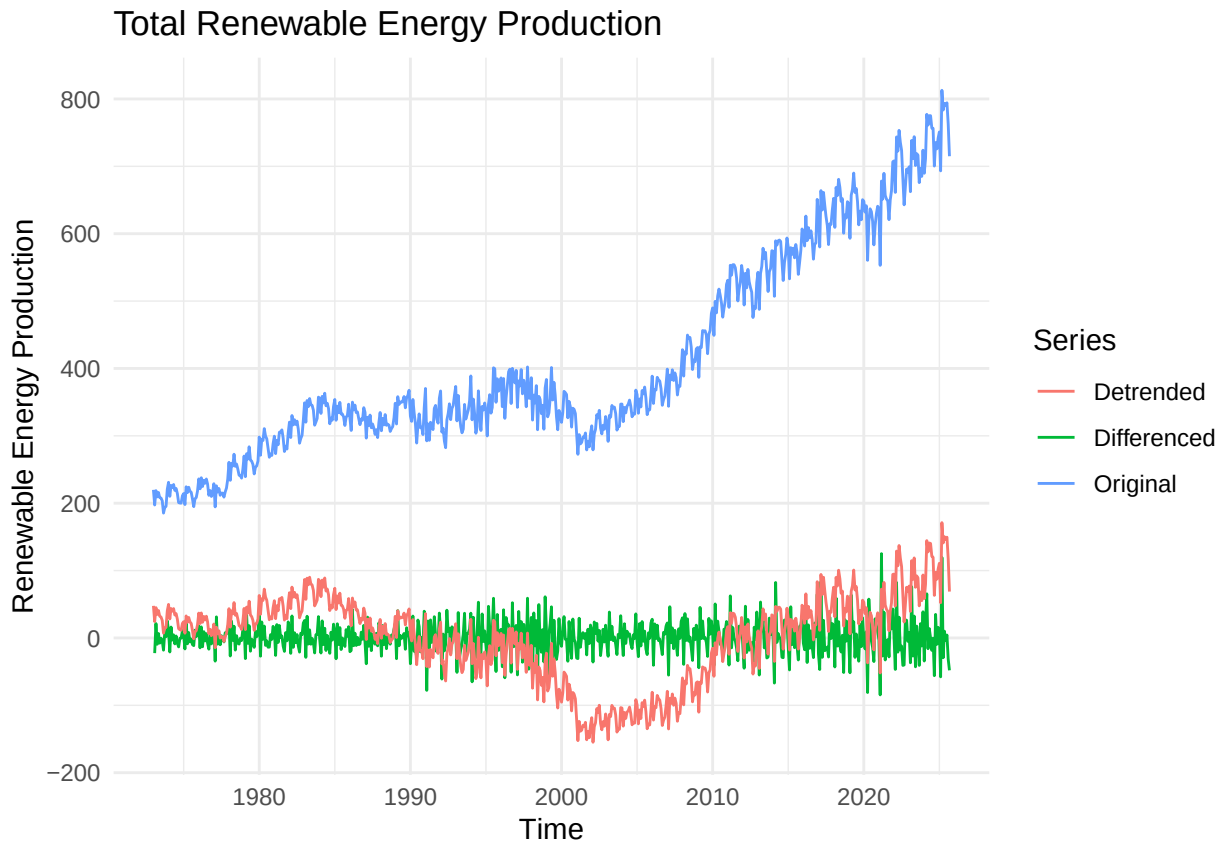
Q3

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 (i.e. "Original", "Differenced", "Detrended lm()"). 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?

```
# Plots for three time series
autoplot(ts_energy_data[,1], series = "Original") +
  autolayer(ts_diff_renew, series = "Differenced") +
  autolayer(ts_renew_detrend, series = "Detrended") +
  labs(
    title = "Total Renewable Energy Production",
    x = "Time",
    y = "Renewable Energy Production",
    color = "Series"
  ) +
  theme_minimal()
```



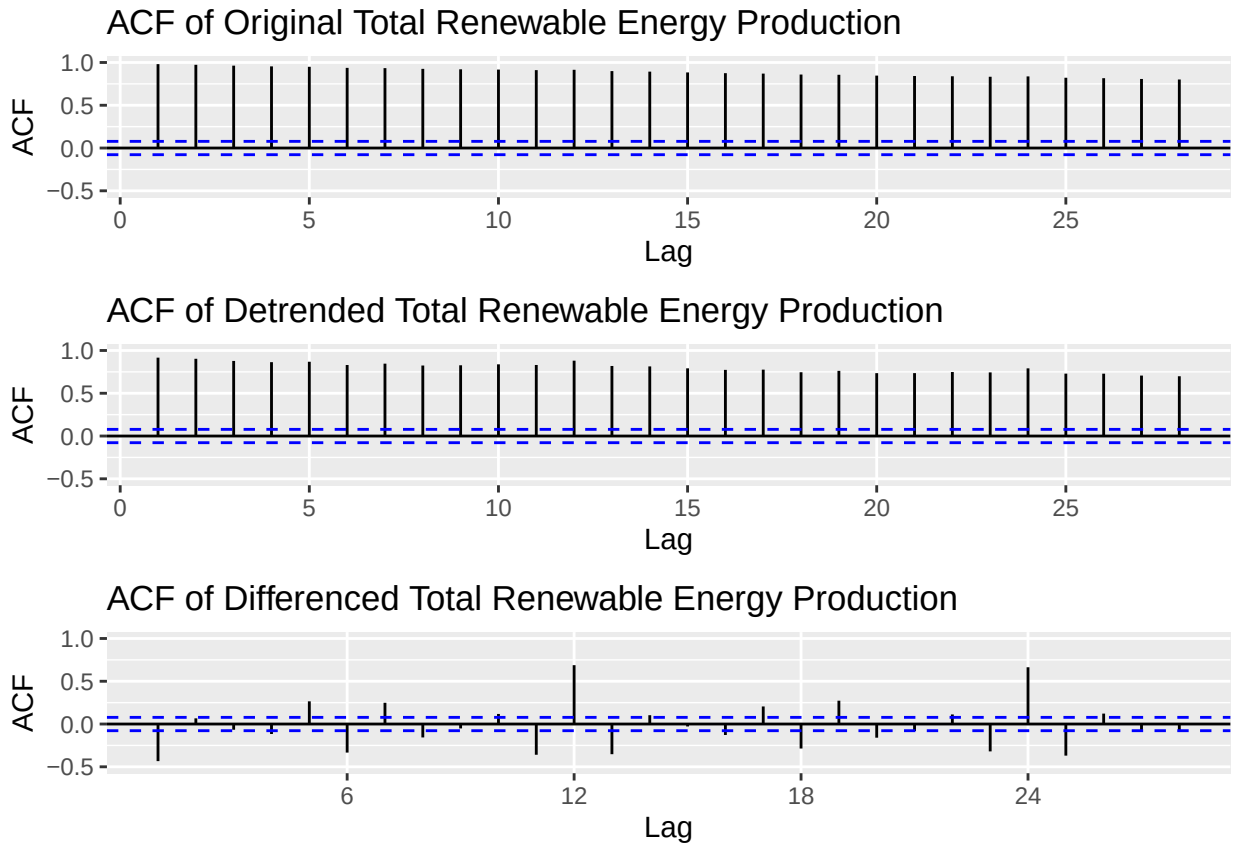
Answer: From the plot, I think differencing is more efficient in removing the trend. The original series shows a strong upward trend, and the detrended series still shows residual long-term structure. However, the differenced series fluctuates around zero.

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?

```
original_acf_plot = ggAcf(energy_select[,2]) +
  ggtitle("ACF of Original Total Renewable Energy Production") +
  ylim(-0.5,1)
detrend_acf_plot = ggAcf(renew_detrend) +
  ggtitle("ACF of Detrended Total Renewable Energy Production") +
  ylim(-0.5,1)
diff_acf_plot = ggAcf(diff_renew) +
  ggtitle("ACF of Differenced Total Renewable Energy Production") +
  ylim(-0.5,1)

plot_grid(original_acf_plot, detrend_acf_plot, diff_acf_plot, ncol = 1)
```



Answer: From the plots, differencing appears to be more efficient. The ACFs of the original series and the detrended series decay slowly, indicating strong persistence driven by the trend. However, the ACF of the differenced series decays rapidly toward zero, meadning that the trend has been removed.

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.

```
# Original Time series
renew_original <- ts_energy_data[,1]

# Seasonal Mann-Kendall
smk_original <- SeasonalMannKendall(renew_original)
smk_original
```

```
## tau = 0.799, 2-sided pvalue =< 2.22e-16
```

```
# ADF Test
adf_original <- adf.test(renew_original)
adf_original
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  renew_original
## Dickey-Fuller = -1.0247, Lag order = 8, p-value = 0.9347
## alternative hypothesis: stationary
```

> Answer: The Seasonal Mann-Kendall test shows tau value of 0.799 and p-value of =< 2.22e-16, which is smaller than 0.05, indicating the trend is statistically significant. Therefore, total renewable energy production shows a significant increasing trend over tiem. The ADF test has a p-value of 0.9347, which is bigger than 0.05. So, we fail to reject the null hypothesis of a unit root. Thus, the series is non-stationary and has a stochastic trend.
> These results matches with the observations in the Q3 plot. The plots in Q3 show very slow decay in ACF, indicating the series is non-stationary.

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().

```
renew_matrix <- matrix(renew_original, nrow = 12, byrow = FALSE)
```

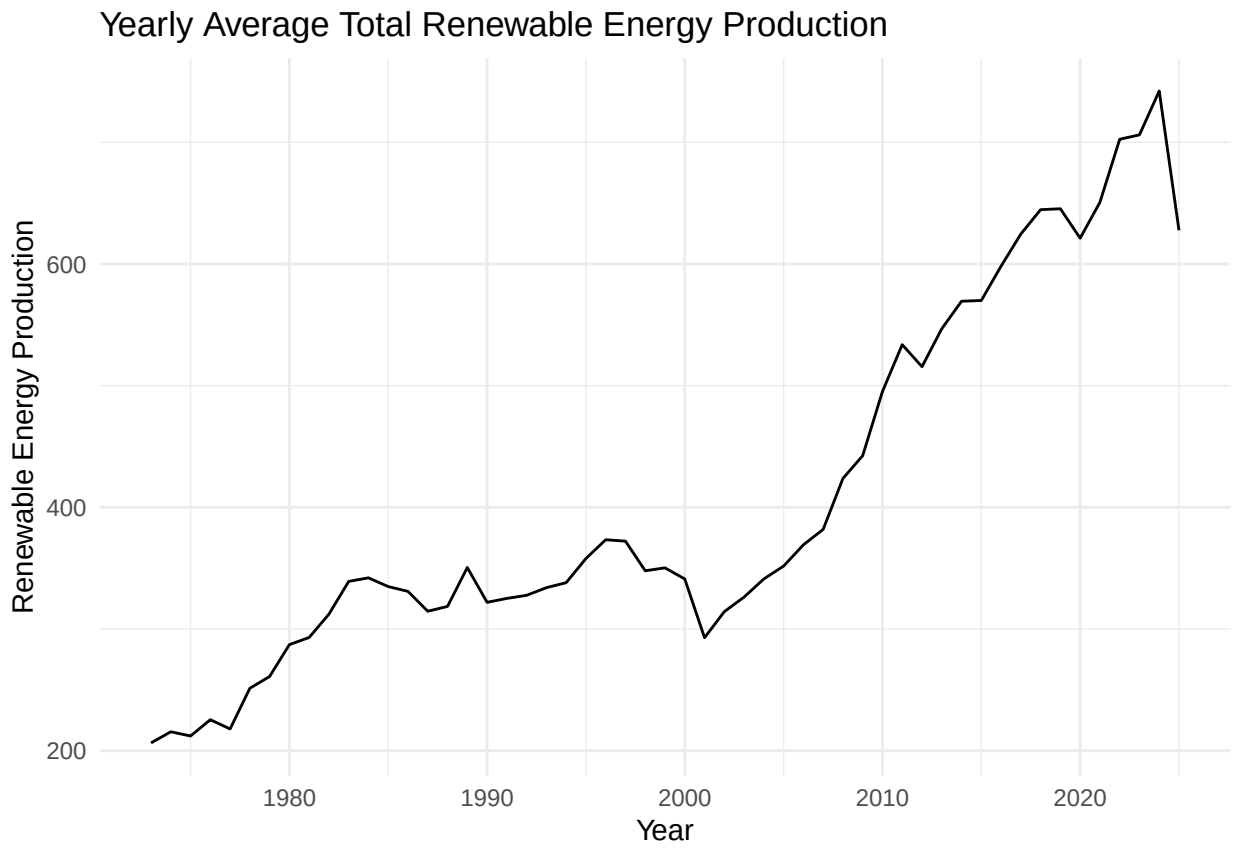
```
## Warning in matrix(renew_original, nrow = 12, byrow = FALSE): data length [633]
## is not a sub-multiple or multiple of the number of rows [12]
```

```r
renew_yearly <- colMeans(renew_matrix, na.rm = TRUE)

start_year <- start(renew_original)[1]

ts_renew_yearly <- ts(renew_yearly, start = start_year, frequency = 1)

autoplot(ts_renew_yearly) +
  labs(
    title = "Yearly Average Total Renewable Energy Production",
    x = "Year",
    y = "Renewable Energy Production"
  ) +
  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 Q5?

```r
# Mann-Kendall Test
mk <- MannKendall(ts_renew_yearly)
mk
```

```
## tau = 0.816, 2-sided pvalue =< 2.22e-16
```

```r
# Spearman Correlation Rank Test
t_year <- 1:length(ts_renew_yearly)

spearman <- cor.test(ts_renew_yearly, t_year, method = "spearman")
spearman
```

```
##
##  Spearman's rank correlation rho
##
## data:  ts_renew_yearly and t_year
## S = 1898, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9234801
```

```r
# ADF Test
adf <- adf.test(ts_renew_yearly)
adf
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_renew_yearly
## Dickey-Fuller = -1.6789, Lag order = 3, p-value = 0.7037
## alternative hypothesis: stationary
```

Answer: The results form the tests are consistent with the results from Q5. The Mann-Kendall test shows a tau value of 0.816 and p-value of =< 2.22e-16, which is smaller than 0.05, indicating the upward trend is statistically significant. Therefore, total renewable energy production shows a significant increasing trend over tiem.
The ADF test has a p-value of 0.7037, which is greater than 0.05. So, we fail to reject the null hypothesis of a unit root. Thus, the series is non-stationary and has a stochastic trend.
The Spearman's test has a rho of 0.9234801 and a p-value of < 2.2e-16, smaller than 0.05, indicating statistically significant positive monotonic relationship between time and total renewable energy production.