ENV 790.30 - Time Series Analysis for Energy Data | Spring 2023

Assignment 4 - Due date 02/17/23 (New Due Date: 2/20/2023)

Katherine Burley

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_Sp23.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(xlsx)
library(readxl)
library(ggplot2)
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(tseries)
library(Kendall)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(lubridate)
```

```
##
```

```
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(formatR)

# Set option for text to wrap in PDF output
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
```

Questions

Consider the same data you used for A3 from the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumption The data comes from the US Energy Information and Administration and corresponds to the December 2022 Monthly Energy Review. For this assignment you will work only with the column "Total Renewable Energy Production".

```
# Importing data set - using xlsx package
energy_data <- as.data.frame(read_excel(path = "../Data/Table_10.1_Renewable_Energy_Production_and_Consu
    skip = 12, sheet = "Monthly Data", col_names = FALSE))
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
```

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

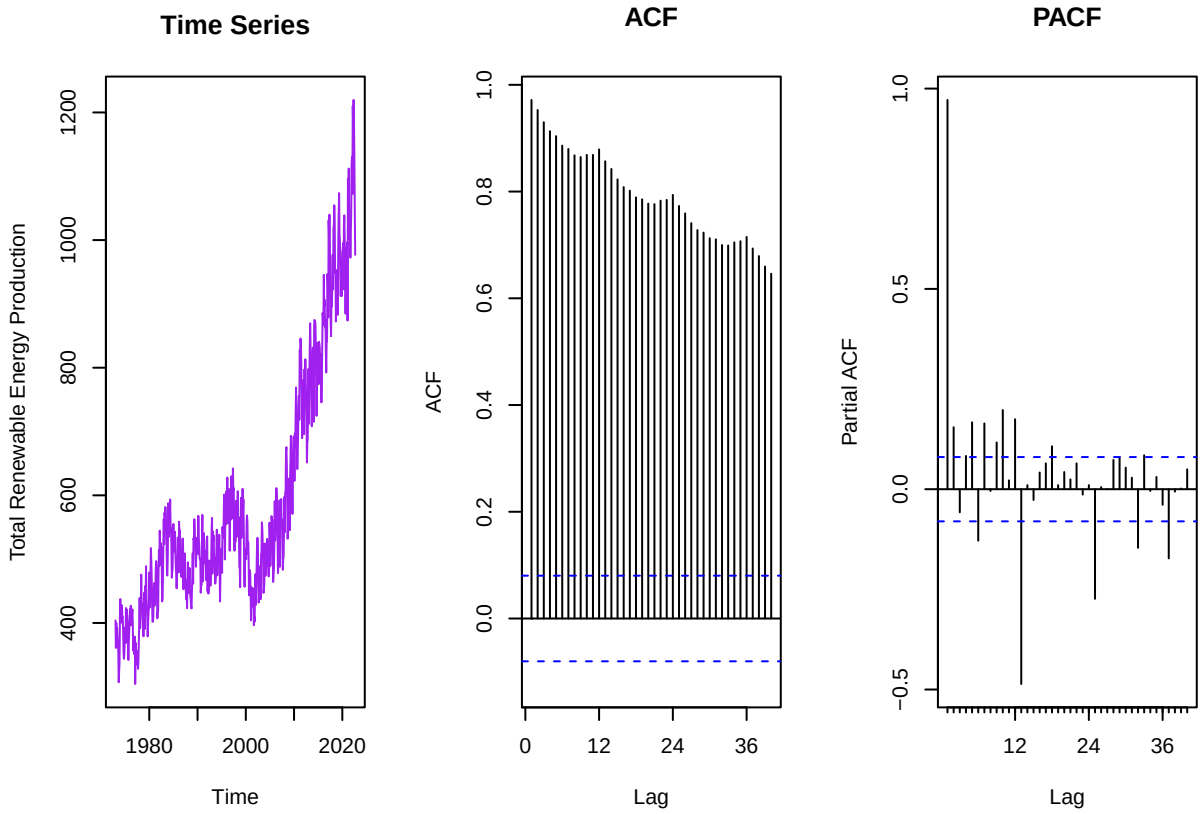
```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
```

```
## * `` -> `...14`
```

```
colnames(energy_data) <- read_col_names

# Subset to Columns We Need
energy_data <- energy_data %>%
    select(c("Month", "Total Renewable Energy Production"))
head(energy_data)
```

```
##        Month Total Renewable Energy Production
## 1 1973-01-01                            403.981
## 2 1973-02-01                            360.900
## 3 1973-03-01                            400.161
## 4 1973-04-01                            380.470
## 5 1973-05-01                            392.141
## 6 1973-06-01                            377.232
```

```
# Create time series object
ts_energy <- ts(energy_data, start = 1973, frequency = 12)  # Monthly data, freq=12

# Plot
par(mfrow = c(1, 3))  #place plot side by side
plot(ts_energy[, 2], main = "Time Series", ylab = "Total Renewable Energy Production",
    col = "purple")
Acf(ts_energy[, 2], lag.max = 40, main = "ACF")
Pacf(ts_energy[, 2], lag.max = 40, main = "PACF")
```



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?

```
diff_re <- diff(energy_data$`Total Renewable Energy Production`,
    lag = 1, difference = 1)
diff_re <- append(NA, diff_re)

energy_data <- energy_data %>%
    bind_cols(diff_re) %>%
    rename(diff_re = "...3")
```

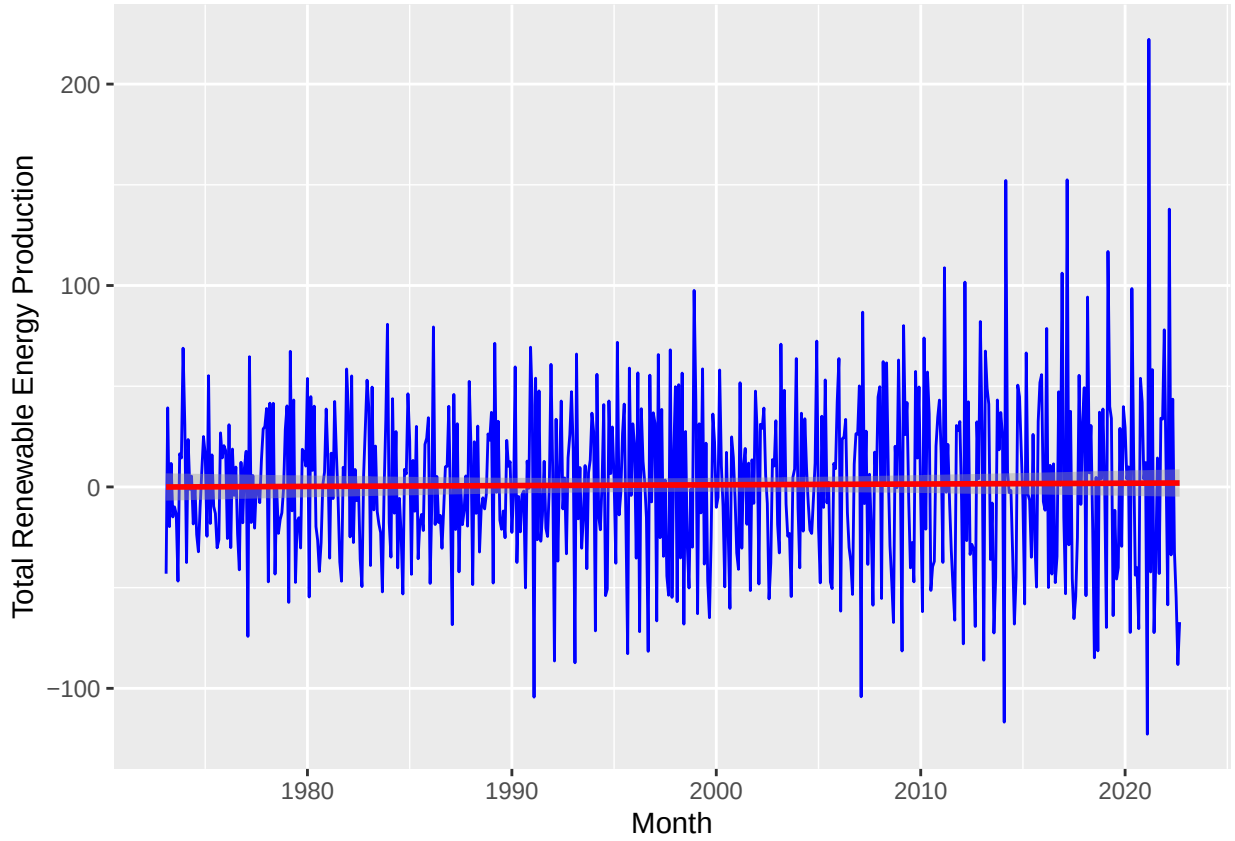
```
## New names:
## * `` -> `...3`
```

```
# Plot
ggplot(energy_data, aes(x = Month, y = diff_re)) + geom_line(color = "blue") +
    ylab("Total Renewable Energy Production") + geom_smooth(color = "red",
    method = "lm")
```

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

```
## Warning: Removed 1 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 1 row containing missing values (`geom_line()`).
```



Differencing the series appears to have removed the trend, with only a very slight positive trend over the time frame.

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

```
nobs <- nrow(energy_data)
t <- c(1:nobs)

re_trend <- lm(energy_data$`Total Renewable Energy Production` ~
    t)
summary(re_trend)
```

```
##
## Call:
## lm(formula = energy_data$`Total Renewable Energy Production` ~
##     t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -238.75  -61.85    8.59   64.48  352.27
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 312.2475     8.4902   36.78   <2e-16 ***
## t             0.9362     0.0246   38.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103.6 on 595 degrees of freedom
## Multiple R-squared:  0.7088, Adjusted R-squared:  0.7083
## F-statistic:  1448 on 1 and 595 DF,  p-value: < 2.2e-16
```

```
detrend_re <- energy_data$`Total Renewable Energy Production` -
    (as.numeric(re_trend$coefficients[1]) + as.numeric(re_trend$coefficients[2]) *
        t)
```

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.

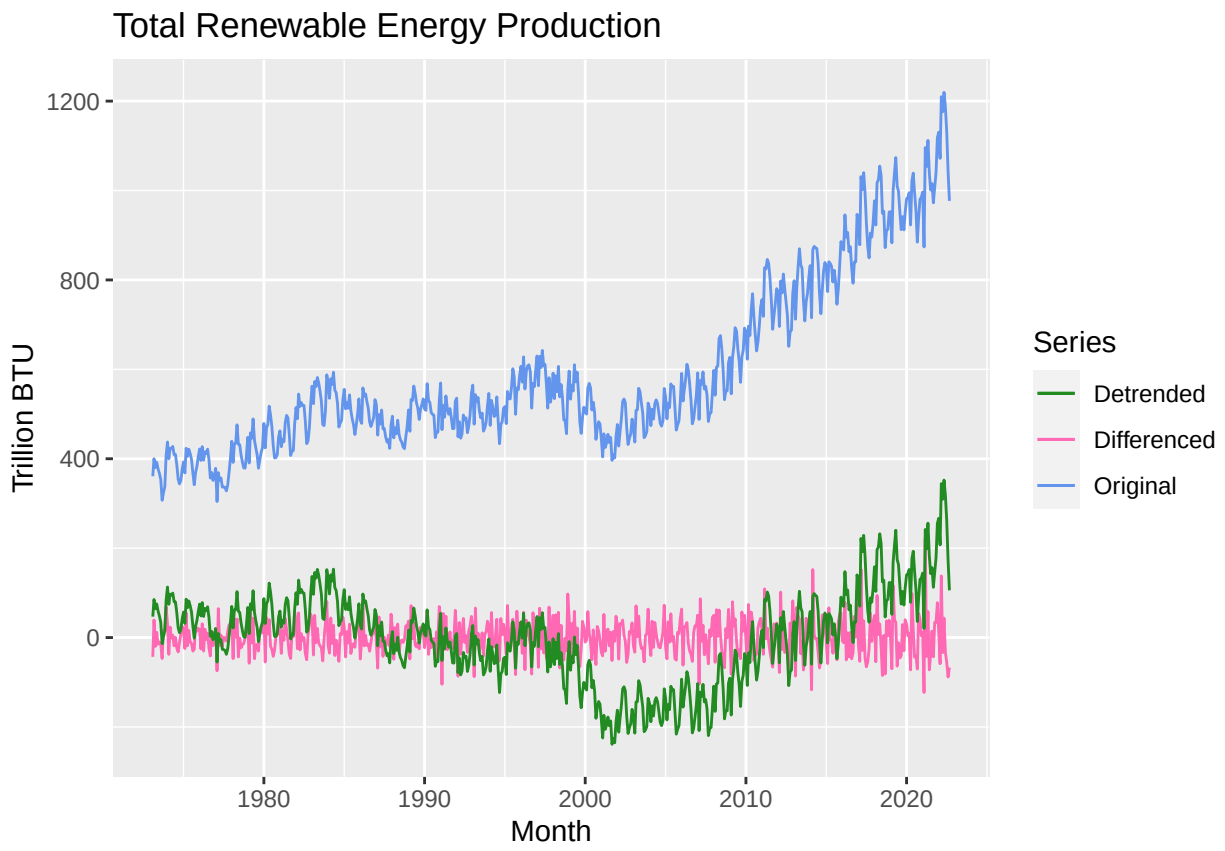
```
# Data frame - remember to not include January 1973 NOTE: I
# had already added an NA observation to the diff_re series
# earlier to make them the same length, but will remove it
# below:

# 1. Remove first observation from the detrended series
detrend_re <- detrend_re[-1]

# 2. Remove 1st observation of Month, the original series,
# and the NA that I added earlier on the differenced series
energy_data_df <- energy_data[-1, ]

# 3. Bind detrended series on to the dataframe, so that
# original, differenced, and detrended series are together
energy_data_df <- energy_data_df %>%
    bind_cols(detrend_re) %>%
    rename(detrend_re = "...4")
```

```
## New names:
## * `` -> `...4`
```

Q4

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

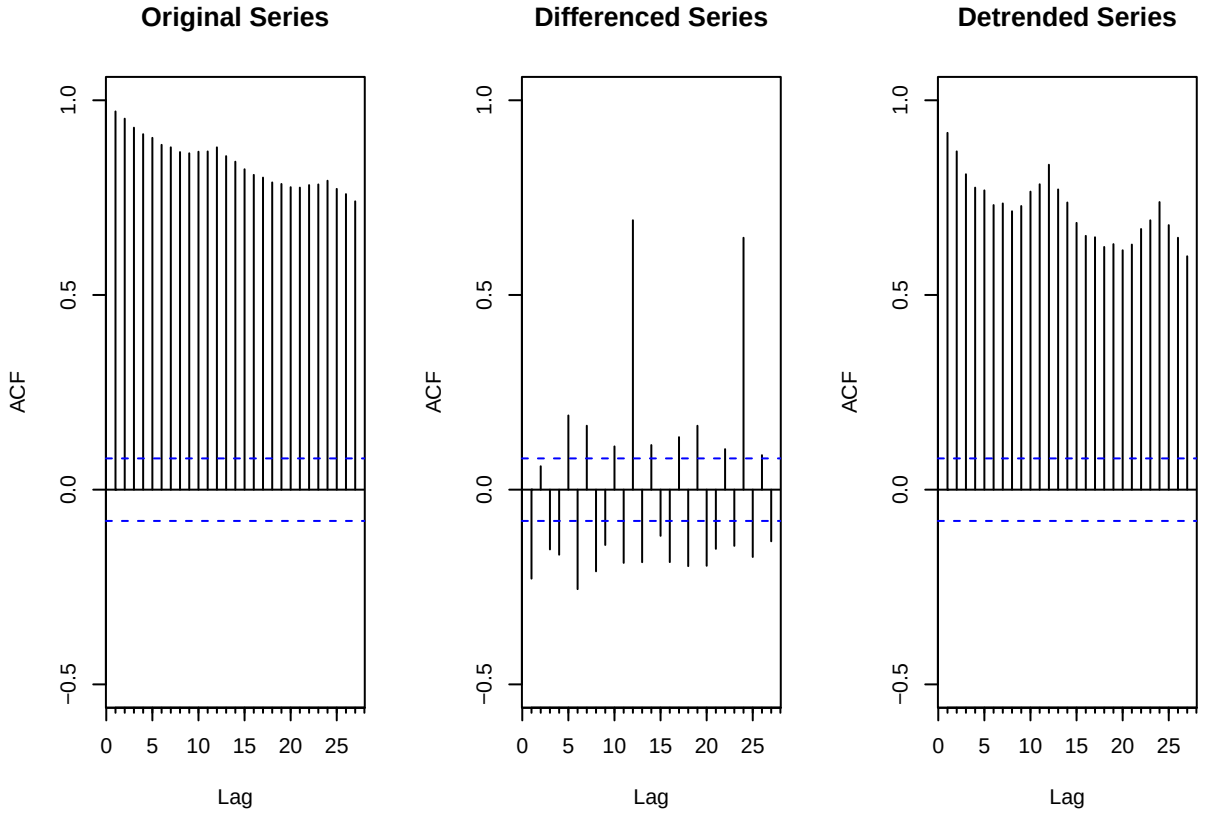
```
# Use ggplot
ggplot(energy_data_df) + geom_line(aes(x = Month, y = `Total Renewable Energy Production`,
    color = "Original")) + geom_line(aes(x = Month, y = diff_re,
    color = "Differenced")) + geom_line(aes(x = Month, y = detrend_re,
    color = "Detrended")) + scale_color_manual(name = "Series",
    values = c(Original = "cornflowerblue", Differenced = "hotpink",
        Detrended = "forestgreen")) + ylab("Trillion BTU") +
    ggtitle("Total Renewable Energy Production")
```



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?

```
# Compare ACFs
par(mfrow = c(1, 3))  #place plot side by side
Acf(energy_data_df$`Total Renewable Energy Production`, main = "Original Series",
    ylim = c(-0.5, 1))
Acf(energy_data_df$diff_re, main = "Differenced Series", ylim = c(-0.5,
    1))
Acf(energy_data_df$detrend_re, main = "Detrended Series", ylim = c(-0.5,
    1))
```



Differencing seems to have done a better trend of eliminating the trend than detrending with the regression coefficients. The original series and detrended series both seem to have a remaining trend, demonstrated by the high autocorrelation value that has a strong relationship with the number of lags, decreasing gradually as lags increase. Most of the ACF scores for the differenced series are much lower in magnitude, fluctuate randomly from negative to positive, and don't seem to have a distinct relationship with the number of lags.

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.

```
# Seasonal Mann Kendall
SMKtest <- SeasonalMannKendall(ts_energy[, 2])
print("Results for Seasonal Mann Kendall")
```

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

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

```
## Score =  10577 , Var(Score) = 169001
## denominator =  14553
## tau = 0.727, 2-sided pvalue =< 2.22e-16
## NULL
```

We reject the null hypothesis that Total RE Production is stationary - we conclude that there is a trend in Total RE Production.

```r
# Augmented Dickey Fuller
print("Results for ADF test")
```

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

```r
print(adf.test(ts_energy[, 2], alternative = "stationary"))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_energy[, 2]
## Dickey-Fuller = -1.2055, Lag order = 8, p-value = 0.9056
## alternative hypothesis: stationary
```

We cannot reject the null hypothesis that the model has a unit root and demonstrates a stochastic trend, suggesting that the series is not stationary (alternative hypothesis).

These tests agree in suggesting that the original series has a non-stationary trend.

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.

```r
# Reshape the dataframe
re_matrix <- matrix(ts_energy[, 2], byrow = FALSE, nrow = 12)
```

```
## Warning in matrix(ts_energy[, 2], byrow = FALSE, nrow = 12): data length [597]
## is not a sub-multiple or multiple of the number of rows [12]
```

```r
# Take column means to get annual data and remove seasonal
# trend
re_annual <- colMeans(re_matrix)
ts_energy_annual <- ts(re_annual, start = 1973, frequency = 1)

my_date <- energy_data_df[, 1]
head(my_date)
```

```
## [1] "1973-02-01 UTC" "1973-03-01 UTC" "1973-04-01 UTC" "1973-05-01 UTC"
## [5] "1973-06-01 UTC" "1973-07-01 UTC"
```

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

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?

```r
# 1. Seasonal Mann Kendall
SMKtest_annual <- MannKendall(ts_energy_annual)
print("Results for Mann Kendall on Annual Data")
```

```
## [1] "Results for Mann Kendall on Annual Data"
```

```r
print(summary(SMKtest_annual))
```

```
## Score =  905 , Var(Score) = 14291.67
```

```
## denominator =  1225
## tau = 0.739, 2-sided pvalue =< 2.22e-16
## NULL
```

```
# 2. Spearman Correlation Rank Test
print("Results for Spearman Correlation Rank Test on Annual Data")
```

```
## [1] "Results for Spearman Correlation Rank Test on Annual Data"
```

```
sp_rho <- cor.test(ts_energy_annual, my_year, method = "spearman")
print(sp_rho)
```

```
##
##  Spearman's rank correlation rho
##
## data:  ts_energy_annual and my_year
## S = 2568, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8766867
```

```
# 3. ADF Test
print("Results for ADF Test on Annual Data")
```

```
## [1] "Results for ADF Test on Annual Data"
```

```
print(adf.test(ts_energy_annual, alternative = "stationary"))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_energy_annual
## Dickey-Fuller = -1.6251, Lag order = 3, p-value = 0.7247
## alternative hypothesis: stationary
```

1. Reject the null hypothesis that annual RE production is stationary, suggesting there is a trend
2. Reject the null hypothesis that annual RE production is stationary (rho == 0), suggesting there is a trend
3. We cannot reject the null hypothesis that the series contains a unit root and displays a trend.

The three tests on the annual RE data agree that renewable energy production displays a non-stationary trend, which also agrees with the results from the tests in question 6 on the original monthly data.