ENV 790.30 - Time Series Analysis for Energy Data | Spring 2022

Assignment 4 - Due date 02/17/22

Aasha Reddy

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 project open the first thing you will do is change "Student Name" on line 3 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. Rename the pdf file such that it includes your first and last name (e.g., "LuanaLima_TSA_A04_Sp21.Rmd"). 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(lubridate)
```

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

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

```
library(ggplot2)
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 4.1.2
```

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

```
library(Kendall)
```

```
## Warning: package 'Kendall' was built under R version 4.1.2
```

```
library(tseries)
```

```
## Warning: package 'tseries' was built under R version 4.1.2
```

```
library(outliers)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x lubridate::as.difftime() masks base::as.difftime()
## x lubridate::date()        masks base::date()
## x dplyr::filter()          masks stats::filter()
## x lubridate::intersect()   masks base::intersect()
## x dplyr::lag()             masks stats::lag()
## x lubridate::setdiff()     masks base::setdiff()
## x lubridate::union()       masks base::union()
```

```
library(readxl)
```

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 January 2021 Monthly Energy Review. For this assignment you will work only with the column "Total Renewable Energy Production".

```
#Importing data set - using xlsx package
repc_raw <- read_excel("/Users/Aasha Reddy/Documents/Statistics - Duke University/2022 Spring/Time Seri
repc_raw <- repc_raw[-1,]
```

```
# Clean data
repc_raw <- repc_raw %>%
  select(`Total Biomass Energy Production`,
         `Total Renewable Energy Production`,
         `Hydroelectric Power Consumption`,
         Month)

# change variables to numeric
repc_raw <- repc_raw %>%
  mutate(`Total Biomass Energy Production` = as.numeric(`Total Biomass Energy Production`),
         `Total Renewable Energy Production` = as.numeric(`Total Renewable Energy Production`),
         `Hydroelectric Power Consumption` = as.numeric(`Hydroelectric Power Consumption`))

# Change Month column to date
repc_raw <- repc_raw %>%
  mutate(Month = ymd(Month))

# transform data into time series object
repc <- ts(data = repc_raw %>% select(-Month), start = c(1973, 1), frequency = 12)
```

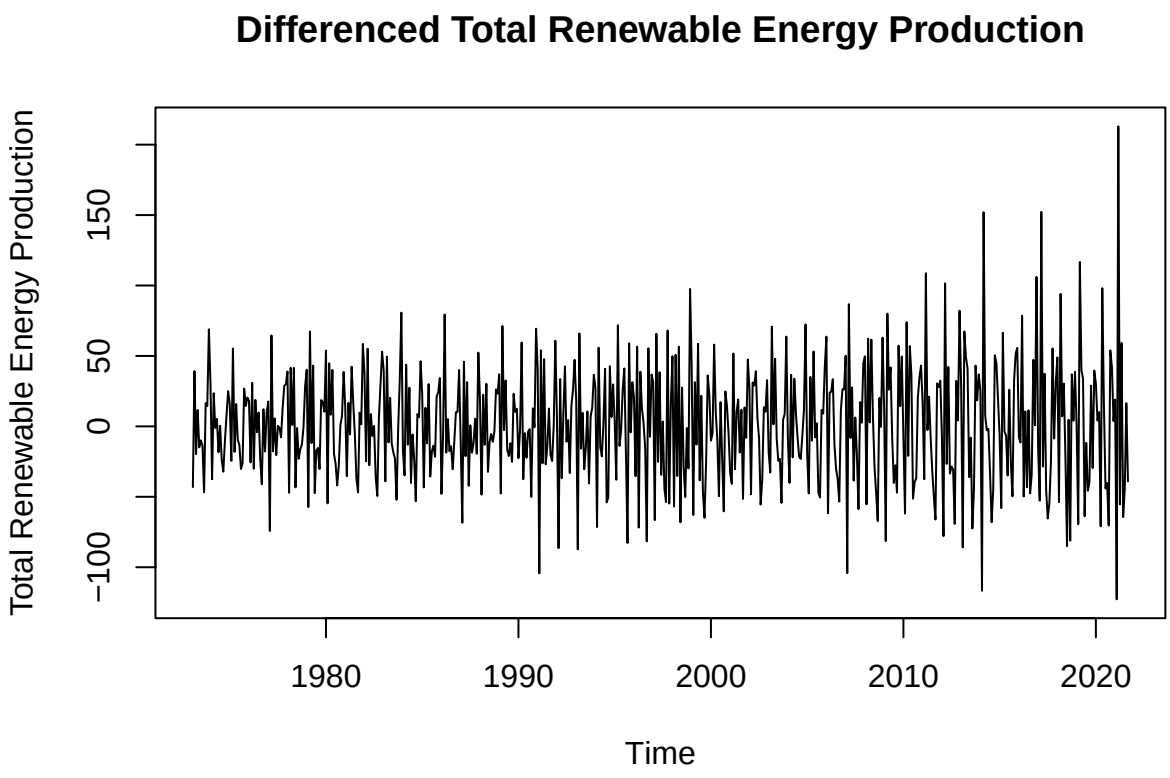
Stochastic Trend and Stationarity Tests

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?

```r
re_diff <- diff(repc[, 2], lag = 1, differences = 1)

plot(re_diff,
     main = "Differenced Total Renewable Energy Production",
     ylab = "Total Renewable Energy Production")
```



After differencing, we can see that the series still seems to have a very very small trend. It looks like the trend is very slightly positively increasing. However, I would saythat overall the series does not still seem to have a trend.

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

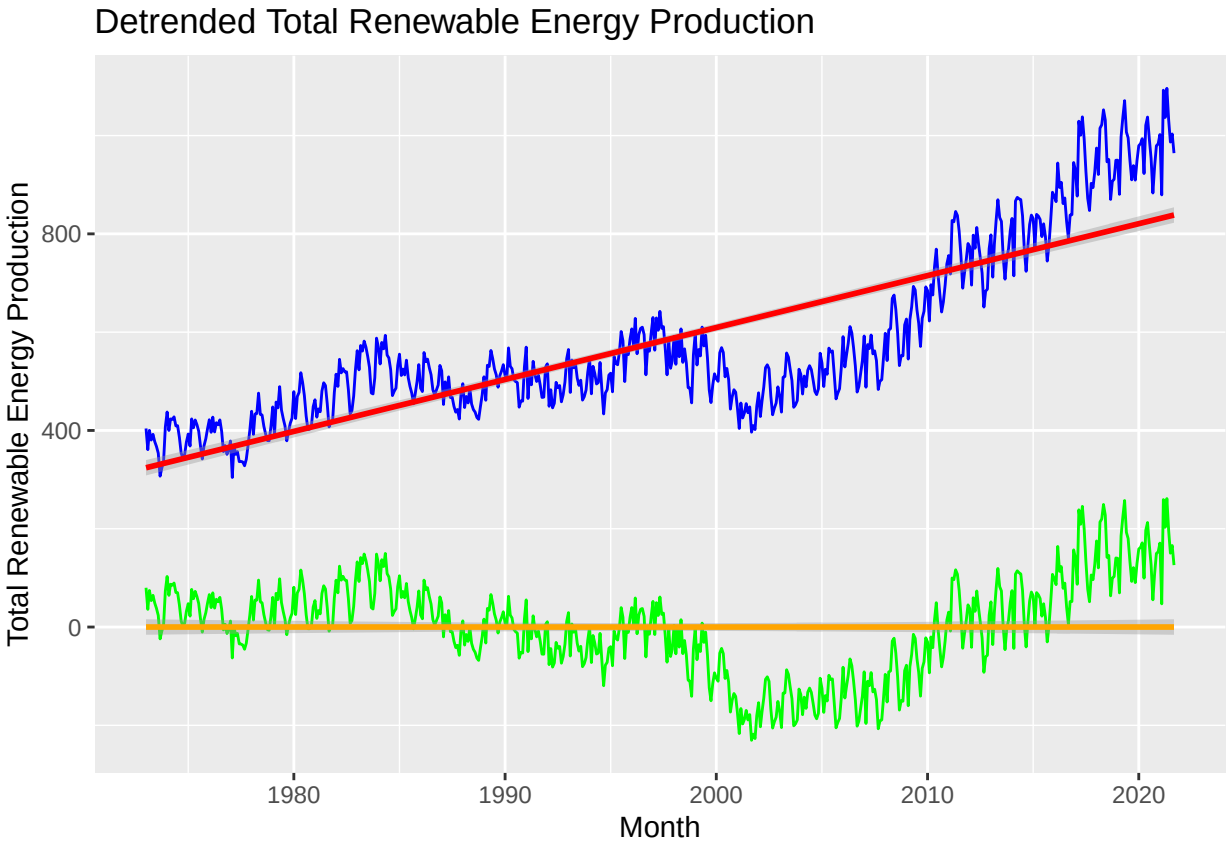
```r
# Detrend series (from A3)

#Create vector t
t <- c(1:nrow(repc))
re_linear <- lm(repc[,2]~t)
summary(re_linear)
```

```
##
## Call:
```

```
## lm(formula = repc[, 2] ~ t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -230.488  -57.869    5.595   62.090  261.349
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 323.18243    8.02555   40.27   <2e-16 ***
## t             0.88051    0.02373   37.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 96.93 on 583 degrees of freedom
## Multiple R-squared:  0.7025, Adjusted R-squared:  0.702
## F-statistic:  1377 on 1 and 583 DF,  p-value: < 2.2e-16
```

```
# save coefficients
re_linear_coefs = coef(re_linear)

beta0=as.numeric(re_linear_coefs[1])
beta1=as.numeric(re_linear_coefs[2])

# detrend
re_detrend <- repc_raw$`Total Renewable Energy Production`-(beta0+beta1*t)


#Understanding what we did
ggplot(repc_raw, aes(x = Month, y = `Total Renewable Energy Production`)) +
            geom_line(color="blue") +
            geom_smooth(color="red",method="lm") +
            geom_line(aes(y=re_detrend), col="green") +
            geom_smooth(aes(y=re_detrend),color="orange",method="lm") +
  labs(title = "Detrended Total Renewable Energy Production")
```

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



We can see that the differenced series for Total Renewable Energy Production looks very different from the detrended series. The Differenced series seems to be more stable around 0, while the detrended series (Green) moves around a bit more. However, the detrended series does seem to better maintain the general structure of the original series, while the differenced series does not look very similar to the original series.

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.

```
# differenced renewable energy using raw data
re_diff <- diff(repc_raw$`Total Renewable Energy Production`, lag = 1,
                differences = 1)

# remove first observations for repc_raw
repc_raw <- repc_raw[-1,]

# remove first observation from re_detrend
repc_raw$re_detrend <- re_detrend[-1]

# add re_diff to existing dataset
repc_raw$re_diff <- re_diff

#Data frame - remember to note include January 1973
re_raw <- repc_raw %>%
```

```
  select(Month, `Total Renewable Energy Production`, re_detrend, re_diff) %>%
  rename(re_orig = `Total Renewable Energy Production`)
```

Head of data frame:

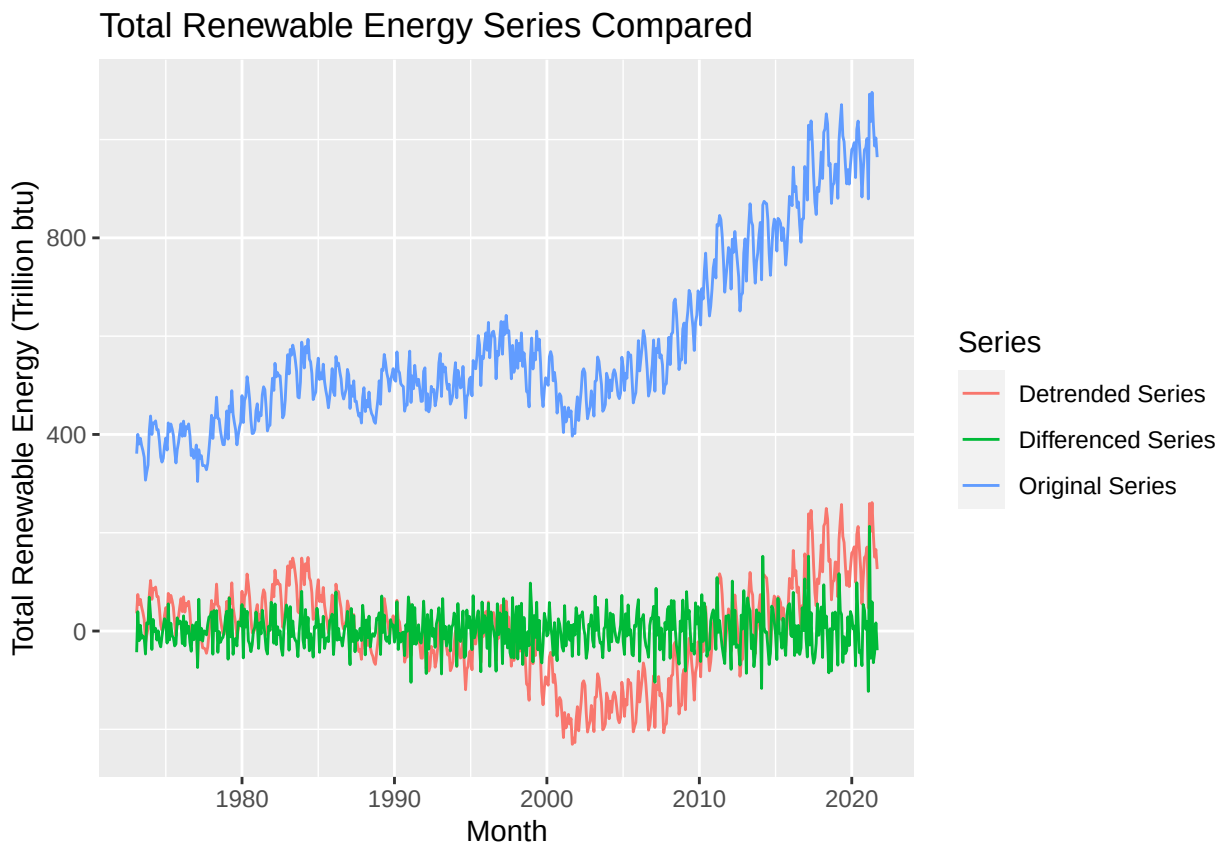
```
head(re_raw)
```

```
## # A tibble: 6 x 4
##   Month      re_orig re_detrend re_diff
##   <date>       <dbl>      <dbl>   <dbl>
## 1 1973-02-01    361.       36.0  -43.1
## 2 1973-03-01    400.       74.3   39.3
## 3 1973-04-01    380.       53.8  -19.7
## 4 1973-05-01    392.       64.6   11.7
## 5 1973-06-01    377.       48.8  -14.9
## 6 1973-07-01    367.       38.0   -9.91
```

Q4

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

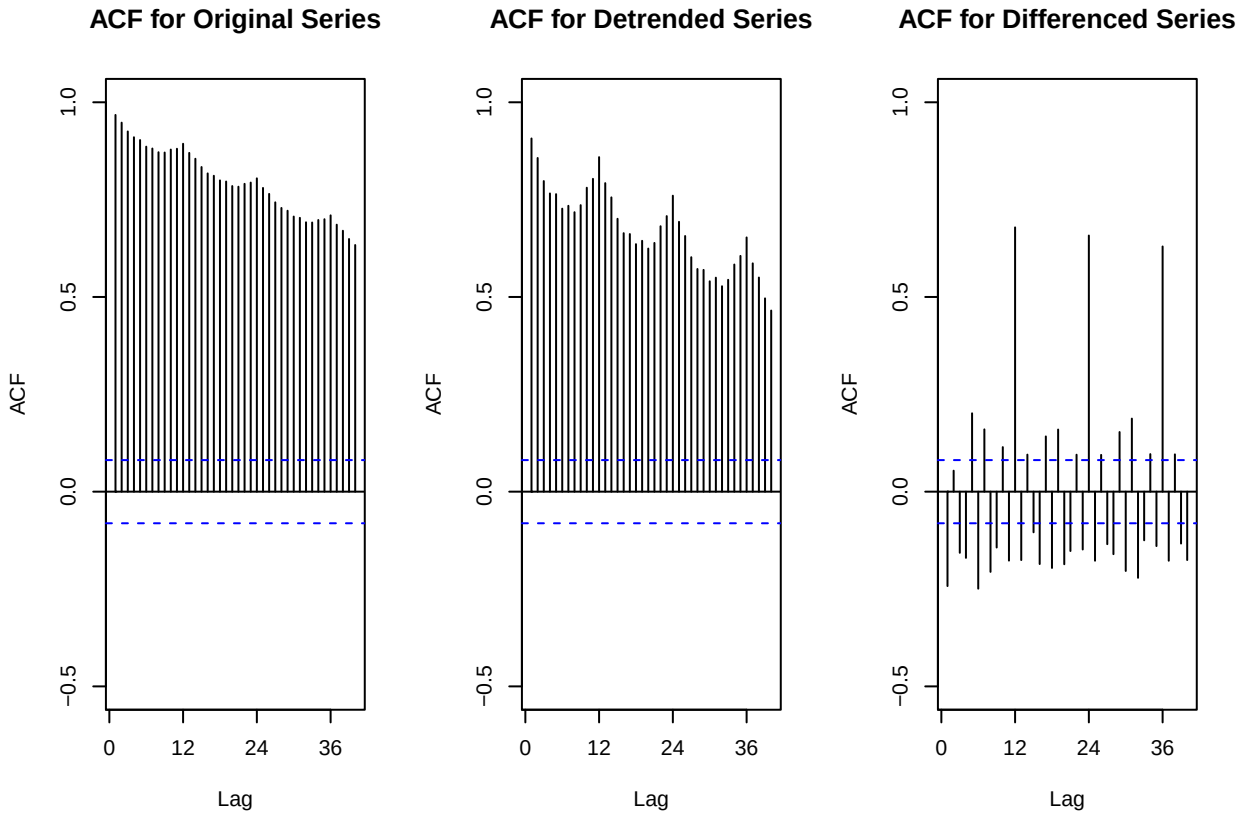
```
# pivot to long form
re_plot <- re_raw %>%
  pivot_longer(2:4, names_to = "Series") %>%
  mutate(Series = case_when(
    Series == "re_orig" ~ "Original Series",
    Series == "re_detrend" ~ "Detrended Series",
    Series == "re_diff" ~ "Differenced Series"
  ))

#Use ggplot
ggplot(data = re_plot, aes(x = Month, y = value, color = Series)) +
  geom_line() +
  labs(title = "Total Renewable Energy Series Compared",
       y = "Total Renewable Energy (Trillion btu)")
```



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

# Divide window into 1 row 3 columns
par(mfrow=c(1, 3))

re_ts <- ts(re_raw,
            frequency = 12,
            start = c(1973, 2))

Acf(re_ts[,2],lag.max=40,
    main="ACF for Original Series",
    ylim = c(-0.5, 1))

Acf(re_ts[,3],lag.max=40,
    main="ACF for Detrended Series",
    ylim = c(-0.5, 1))

Acf(re_ts[,4],lag.max=40,
    main="ACF for Differenced Series",
    ylim = c(-0.5, 1))
```



Based on the above ACF plots, it looks like differencing actually did a better job of betting rid of the trend versus detrending using linear regression.

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 Test on Original Series:

```
# Seasonal Mann-Kendall Test
SMKtest <- SeasonalMannKendall(re_ts[,2])
print("Results for Seasonal Mann Kendall /n")
```

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

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

```
## Score =  9940 , Var(Score) = 158304
## denominator =  13920
## tau = 0.714, 2-sided pvalue =< 2.22e-16
## NULL
```

The Mann-Kendall test allows us to check for a deterministic trend. We can see that our p-value is very low at =< 2.22e-16, which means we reject the null hypothesis that the original series for Total Renewable Energy is stationary.

This provides evidence that the original series for Total Renewable Energy follows a trend.

ADF Test on Original Series:

```
#Null hypothesis is that data has a unit root
print("Results for ADF test/n")
```

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

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

```
##
##  Augmented Dickey-Fuller Test
##
## data:  re_ts[, 2]
## Dickey-Fuller = -1.428, Lag order = 8, p-value = 0.8204
## alternative hypothesis: stationary
```

The ADF test allows us to check for a stochastic trend. We can see that our p-value is very high, at 0.8204, which means that we cannot reject the null hypothesis that the original series contains a unit root.

This suggests that the original series for Total Renewable Energy is stationary and does not have a stochastic trend.

Both of these follow what we saw in question 2. From question 2, the original series does look like it has a deterministic trend as opposed to a stochastic trend because the variance does appear to be relatively constant.

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.

```
re_agg <- re_raw %>%
  mutate(year = year(Month)) %>%
  group_by(year) %>%
  summarize(re_orig = mean(re_orig))

# turn into a time series object
re_agg_ts <- ts(data = re_agg %>% select(-year), start = c(1973), frequency = 1)
```

Head of aggregated data:

```
head(re_agg)
```

```
## # A tibble: 6 x 2
##    year re_orig
##   <dbl>   <dbl>
## 1  1973    364.
## 2  1974    395.
## 3  1975    391.
## 4  1976    394.
## 5  1977    351.
## 6  1978    417.
```

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?

Seasonal Mann-Kendall Test on Original Series:

```
# Seasonal Mann-Kendall Test
SMKtest <- SeasonalMannKendall(re_agg_ts)
print("Results for Seasonal Mann Kendall /n")
```

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

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

```
## Score =  864 , Var(Score) = 13458.67
## denominator =  1176
## tau = 0.735, 2-sided pvalue =9.5035e-14
## NULL
```

The Mann-Kendall test allows us to check for a deterministic trend. We can see that our p-value is very low at 9.5035e-14, which means we reject the null hypothesis that the original series for Total Renewable Energy is stationary.

This provides evidence that the original series for Total Renewable Energy follows a trend.

ADF Test on Original Series:

```
#Null hypothesis is that data has a unit root
print("Results for ADF test/n")
```

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

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

```
##
##  Augmented Dickey-Fuller Test
##
## data:  re_agg_ts
## Dickey-Fuller = -0.84269, Lag order = 3, p-value = 0.9519
## alternative hypothesis: stationary
```

The ADF test allows us to check for a stochastic trend. We can see that our p-value is very high, at 0.9519, which means that we cannot reject the null hypothesis that the original series contains a unit root.

This suggests that the original series for Total Renewable Energy is stationary and does not have a stochastic trend.

Both of these results are in agreement with the non-aggregated results from question 6.