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

Assignment 4 - Due date 02/11/25

Jessalyn Chuang

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(lubridate)
library(ggplot2)
library(forecast)
library(Kendall)
library(tseries)
library(readxl)
library(base)
library(cowplot)
library(tidyverse)
```

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”**.

```
#Importing data set - you may copy your code from A3
energy_data <- read_excel(path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source
#Extract column names from from row 11
read_col_names <- read_excel(path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sou

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

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

```
## # 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>, ...
```

```
#pulling Renewable Energy Production only
renewable <- energy_data[, 5]

#turning into a time series
renewable_ts <- ts(renewable[,1],start=c(1973,1),frequency=12)
```

Stochastic Trend and Stationarity Tests

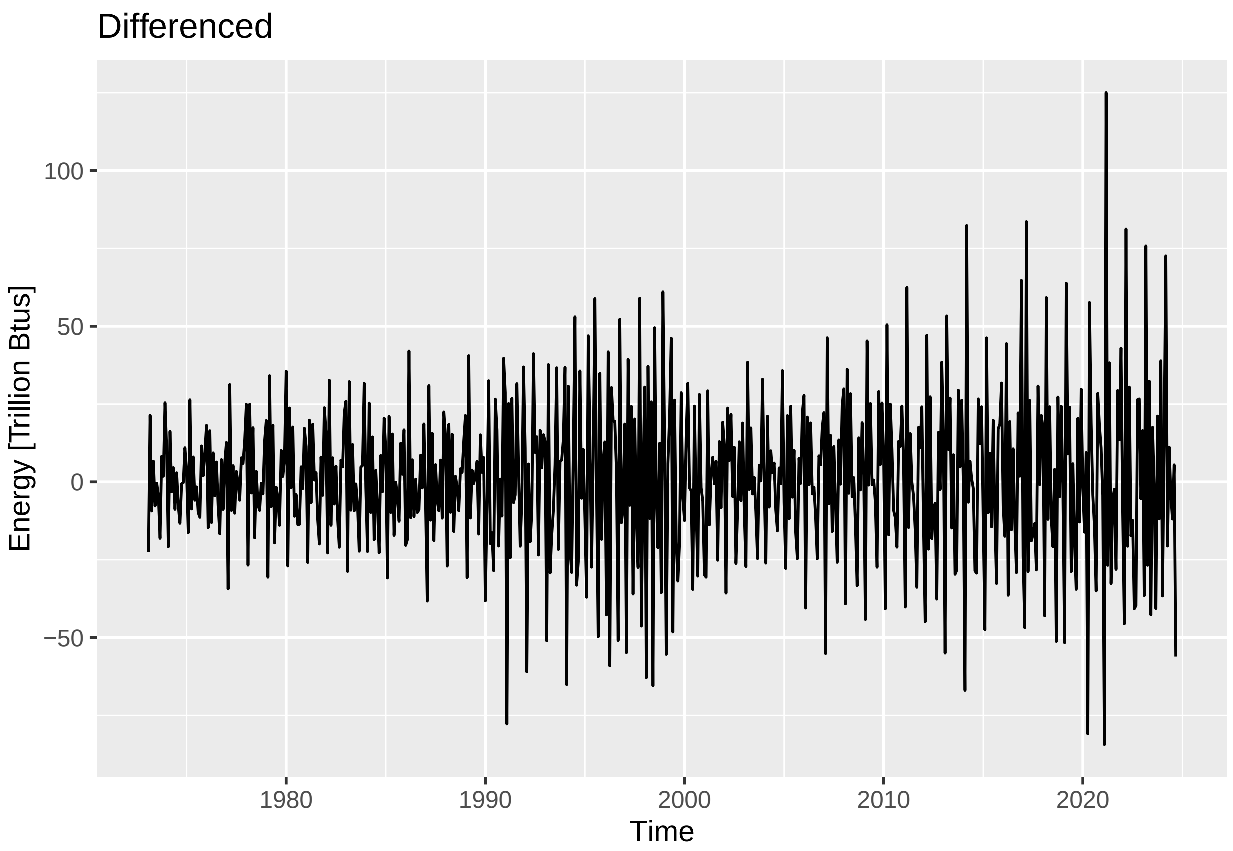
For this part you will work only with the column Total Renewable Energy Production.

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?

```
renewable_diff <- diff(renewable_ts, lag = 1, differences = 1)
autoplot(renewable_diff)+
  ylab("Energy [Trillion Btus]")+
  ggtitle("Differenced")
```



This series does not look to have the trend. Before differencing, there was a clear increasing trend, but differencing at lag 1 helped to remove this.

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.

```
nobs <- nrow(renewable)
col_renewable <- 1
#Create vector t
t <- c(1:nobs)

#Fit a linear trend to Renewable Energy Production
renewable_linear_model <- lm(renewable[[col_renewable]] ~ t)
summary(renewable_linear_model)
```

```
##
## Call:
## lm(formula = renewable[[col_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
```

```
renewable_beta0 <- as.numeric(renewable_linear_model$coefficients[1])
renewable_beta1 <- as.numeric(renewable_linear_model$coefficients[2])

renewable_linear_trend <- renewable_beta0 + renewable_beta1 * t
ts_renewable_linear <- ts(renewable_linear_trend,star=c(1973,1),frequency=12)

detrend_renewable <- renewable[,col_renewable] - renewable_linear_trend
renewable_detrend_ts <- ts(detrend_renewable, start = c(1973,1),frequency = 12)
```

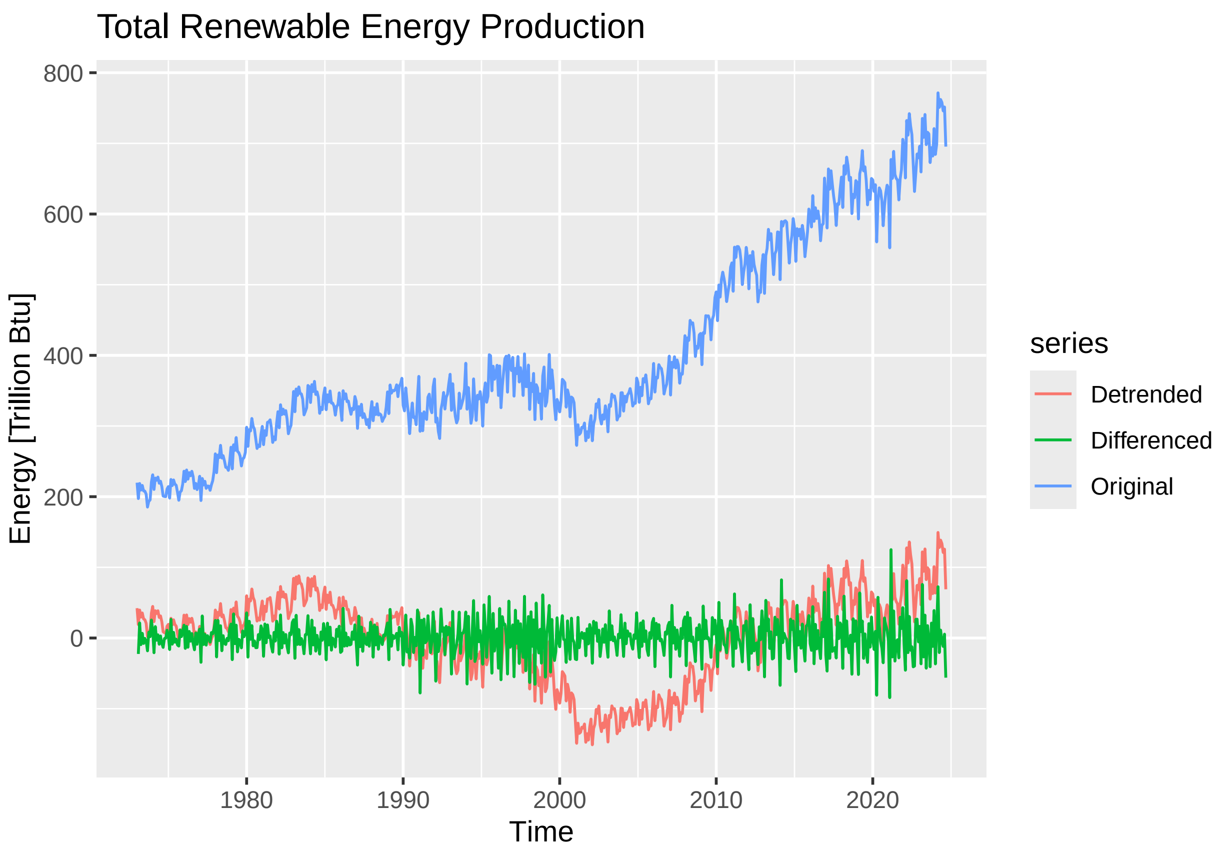
Q3

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?

```
autoplot(renewable_ts[,1], series = "Original") +
  autolayer(renewable_detrend_ts[,1],series="Detrended") +
  autolayer(renewable_diff,series="Differenced")+
  ylab("Energy [Trillion Btu]") +
  ggtitle("Total Renewable Energy Production")
```



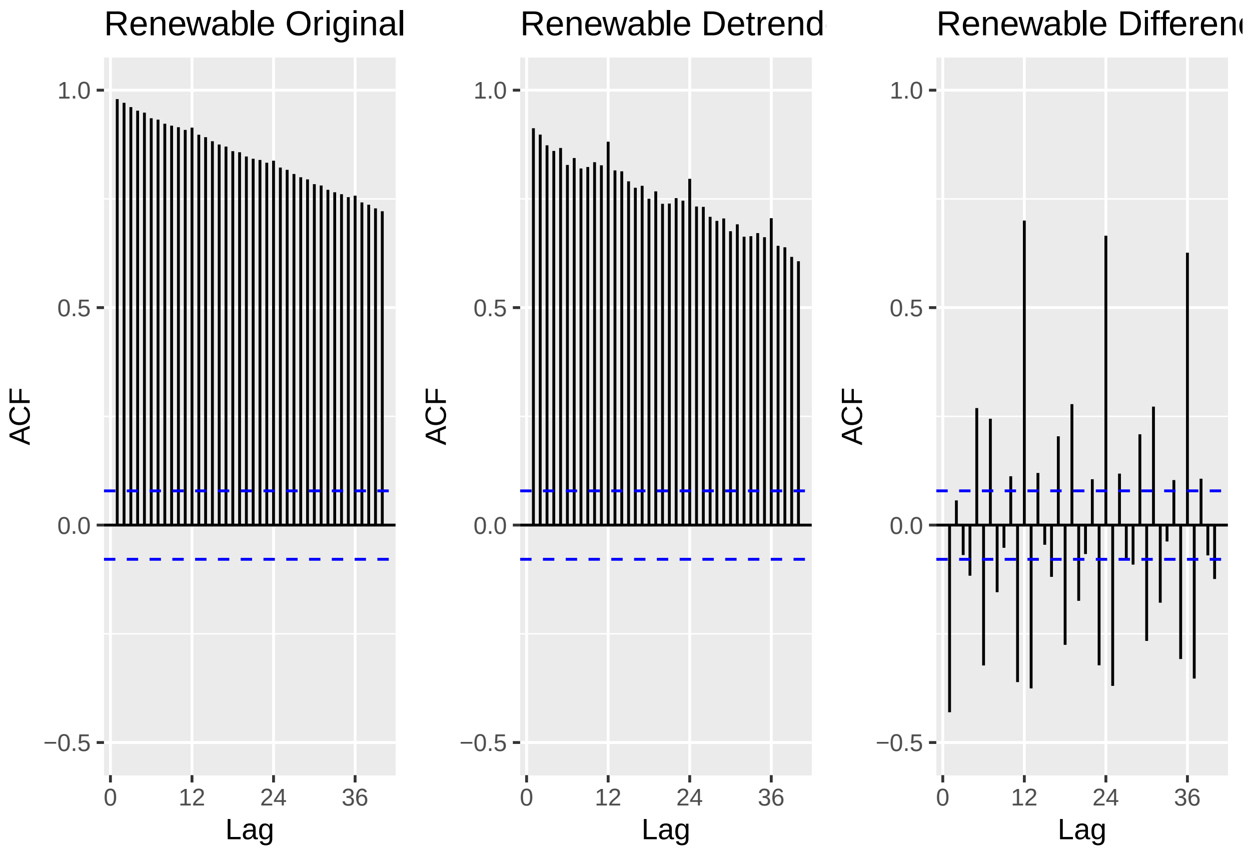
Answer: It looks like the difference method was the most efficient in removing the trend. The plot shows the original renewable energy production data (blue) with a clear upward trend, while the detrended series (red) and differenced series (green) attempt to remove this trend. The detrended series still exhibits long-term variations, indicating incomplete trend removal, whereas the differenced series appears more stationary, with fluctuations centered 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?

```r
plot_grid(
  autoplot(Acf(renewable_ts,lag.max=40,plot=FALSE),
           main="Renewable Original",
           ylim=c(-0.5,1)),
  autoplot(Acf(renewable_detrend_ts,lag.max=40,plot=FALSE),
           main="Renewable Detrended",
           ylim=c(-0.5,1)),
  autoplot(Acf(renewable_diff,lag.max=40,plot=FALSE),
           main="Renewable Differenced",
           ylim=c(-0.5,1)),
  nrow=1,
  ncol=3
)
```

```
## Warning in ggplot2::geom_segment(lineend = "butt", ...): Ignoring unknown parameters: ‘main‘ and ‘yl
## Ignoring unknown parameters: ‘main‘ and ‘ylim‘
## Ignoring unknown parameters: ‘main‘ and ‘ylim‘
```



Answer: Differencing was the most efficient method for eliminating the trend. The ACF of the original series shows a strong, slow decay, indicating non-stationarity due to the presence of a linear trend. The ACF of the detrended series remains similar to the original, suggesting that detrending did not fully remove the linear trend. In contrast, the ACF of the differenced series fluctuates around zero with much weaker autocorrelations, confirming that differencing effectively removed the linear trend and made the series more stationary.

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
#Seasonal Mann-Kendall
summary(SeasonalMannKendall(renewable_ts))
```

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

```r
#ADF Test
print("Results for ADF test/n")
```

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

```r
print(adf.test(renewable_ts,alternative = "stationary")) #stationary over stochastic trend
```

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

> Answer: The Seasonal Mann-Kendall test showed p-value being significant. This means that we reject the null hypothesis that the data is i.i.d in favor of the alternative hypothesis that the data follows a trend. For the ADF test, since the p-value was 0.92, we accept the null hypothesis that the data contains a unit root meaning that the series has a stochastic trend. Overall, the time series has a trend, as indicated by the Mann-Kendall test. However, the series is still non-stationary as it has a unit root that makes differencing the data necessary to remove the trend and achieve stationarity.

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().

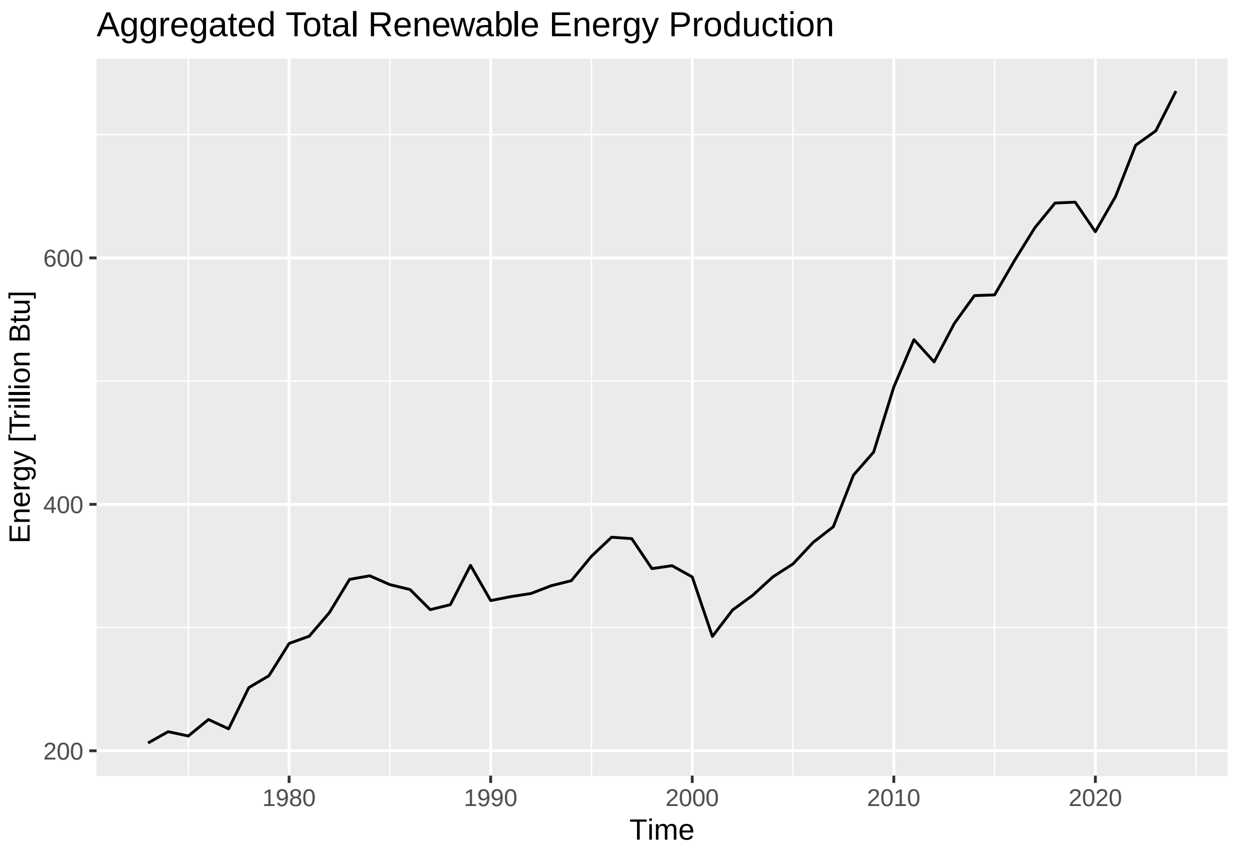
```r
renewable_with_date <- energy_data[, c(1,5)]
renewable_with_date$Month <- ymd(renewable_with_date$Month)

date_matrix <- renewable_with_date %>%
  mutate(Year = year(Month),
         Month = month(Month, label = TRUE))

renewable_matrix <- date_matrix %>%
  group_by(Year, Month) %>%
  summarise(Total_Production = sum(`Total Renewable Energy Production`, na.rm = TRUE), .groups = "drop"
  pivot_wider(names_from = Year, values_from = Total_Production) %>%
  column_to_rownames(var = "Month")
```

```
yearly_series <- colMeans(renewable_matrix, na.rm = TRUE)
yearly_ts <- ts(yearly_series, start = 1973, frequency = 1)

autoplot(yearly_ts) +
  ggtitle("Aggregated Total Renewable Energy Production") +
  ylab("Energy [Trillion Btu]")
```



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?

```
#Mann-Kendall
summary(MannKendall(yearly_ts))
```

```
## Score =  1084 , Var(Score) = 16059.33
## denominator =  1326
## tau = 0.817, 2-sided pvalue =< 2.22e-16
```

```
#Deterministic trend with Spearman Correlation Test
my_year <- c(1973:2024)
print("Results from Spearman Correlation")
```

```
## [1] "Results from Spearman Correlation"
```

```r
print(cor.test(yearly_ts, my_year, method="spearman"))
```

```
##
##  Spearman's rank correlation rho
##
## data:  yearly_ts and my_year
## S = 1852, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9209425
```

```r
#ADF Test
print("Results for ADF test/n")
```

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

```r
print(adf.test(yearly_ts,alternative = "stationary")) #stationary over stochastic trend
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  yearly_ts
## Dickey-Fuller = -0.93521, Lag order = 3, p-value = 0.9399
## alternative hypothesis: stationary
```

> Answer: Yes, the results from the Mann Kendall and ADF tests were similar (reject the null hypothesis for Mann Kendall, meaning that the data follows a trend; reject the null hypothesis for the Spearman Correlation and accept the alternate hypothesis stating that rho is not equal to 0 and that there is a trend; accept the null hypothesis for ADF meaning that the data contains a unit root and has a stochastic trend). The Spearman Correlation test result showed that the correlation coefficient is 0.92, meaning that there is an increasing trend, which is apparent from the plot as well!