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

Assignment 4 - Due date 02/11/25

Lucy Wang

## 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(readxl)
library(openxlsx)
library(tseries)
library(Kendall)
library(cowplot)
library(glue)
library(trend)
```

## 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 - you may copy your code from A3
data_file <- read_excel(path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.x
                        skip = 12,
                        sheet="Monthly Data",col_names=FALSE)
#Extract the column names from row 11
```

```
read_col_names <- read_excel(path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sou
                             skip = 10,n_max = 1,
                             sheet="Monthly Data",col_names=FALSE)

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

#Visualize the first rows of the data set
cleaned_df <- subset(data_file, select = c('Month',
                                           'Total Renewable Energy Production'))
cleaned_df <- as.data.frame(cleaned_df)

#Create time series
ts_data <- ts(cleaned_df[,2],start=c(1973,1),frequency=12)

nobs <- nrow(cleaned_df)
```

## Stochastic Trend and Stationarity Tests

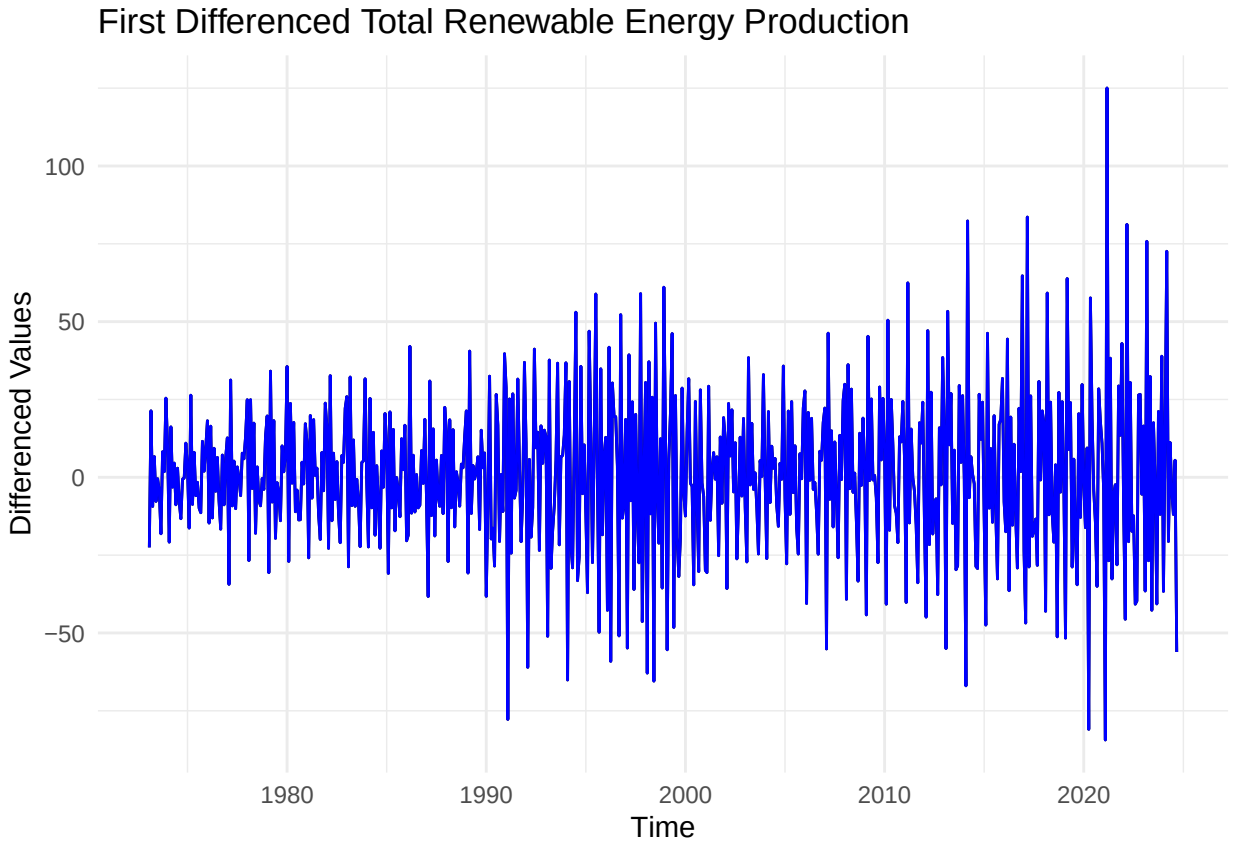
For this part you will work only with the column Total Renewable Energy Production.

### 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?

```
ts_diff <- diff(ts_data, lag = 1, differences = 1)
autoplot(ts_diff)+
  geom_line(color="blue") +
  ylab("Differenced Values") +
  ggtitle("First Differenced Total Renewable Energy Production") +
  theme_minimal()
```




No, the differenced series does not have a trend. The trend is removed after differencing.

### 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.

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

#Fit a linear trend to TS of renewable
Renewable_linear_trend <- lm(cleaned_df[,2] ~ t)

# detrend function
plot_detrend <- function(lm_model, i){
  # assign beta
  beta0 <- as.numeric(lm_model$coefficients[1])
  beta1 <- as.numeric(lm_model$coefficients[2])

  # detrend inflow
  linear_trend <- beta0 + beta1 * t
  ts_linear <- ts(linear_trend,start=c(1973,1),frequency=12)

  detrend_energy <- cleaned_df[,i+1] - linear_trend
  ts_detrend <- ts(detrend_energy, start = c(1973,1),frequency = 12)
```

```
  #Plot
  detrended_plot <- autoplot(ts_detrend,color="green")+
    ggtitle("Detrend Total Renewable Energy Production") +
    theme_minimal()

  detrended_combined_plot <- autoplot(ts_data,color="darkblue")+
    autolayer(ts_detrend,series="Detrended",color="green")+
    autolayer(ts_linear,series="Linear Component",color="red")+
    ggtitle(colnames(ts_data))

  return(list(detrended_data = ts_detrend, plot = detrended_plot))
}

Renewable_detrend_data <- plot_detrend(Renewable_linear_trend,1)$detrended_data
Renewable_detrend_plot <- plot_detrend(Renewable_linear_trend,1)$plot
```

**Q3**

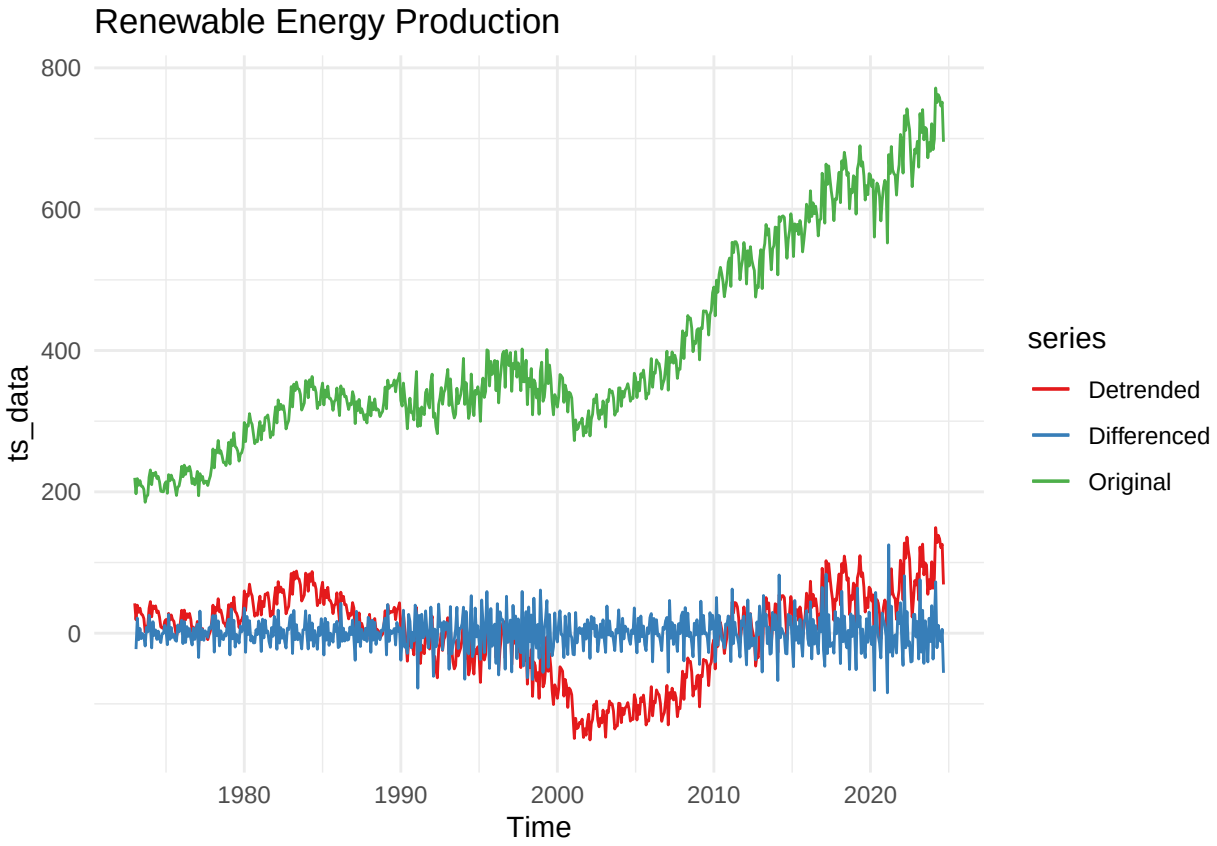
Now let's compare the differenced series with the detrended series you calculated on A3. In other words, for the "Total Renewable Energy Production" compare the differenced series from Q1 with the series you detrended in Q2 using linear regression.

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?

```
combined_plot <- autoplot(ts_data, series = "Original", aes(color=group))+
  autolayer(Renewable_detrend_data,series="Detrended")+
  autolayer(ts_diff,series="Differenced")+
  scale_color_brewer(palette = "Set1") +
  xlab("Time")+
  ggtitle("Renewable Energy Production")+
  theme_minimal()

combined_plot
```




Answer: Differencing the series with function diff() seems to be more efficient in removing the trend as the values fluctuate around zero without an observable upward or downward trend. The original series presents an obvious increasing trend. Though detrend series eliminates partial upward trend, there are still minor downward and upward trends observed. The differenced series presents nearly no trend at all.

**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?

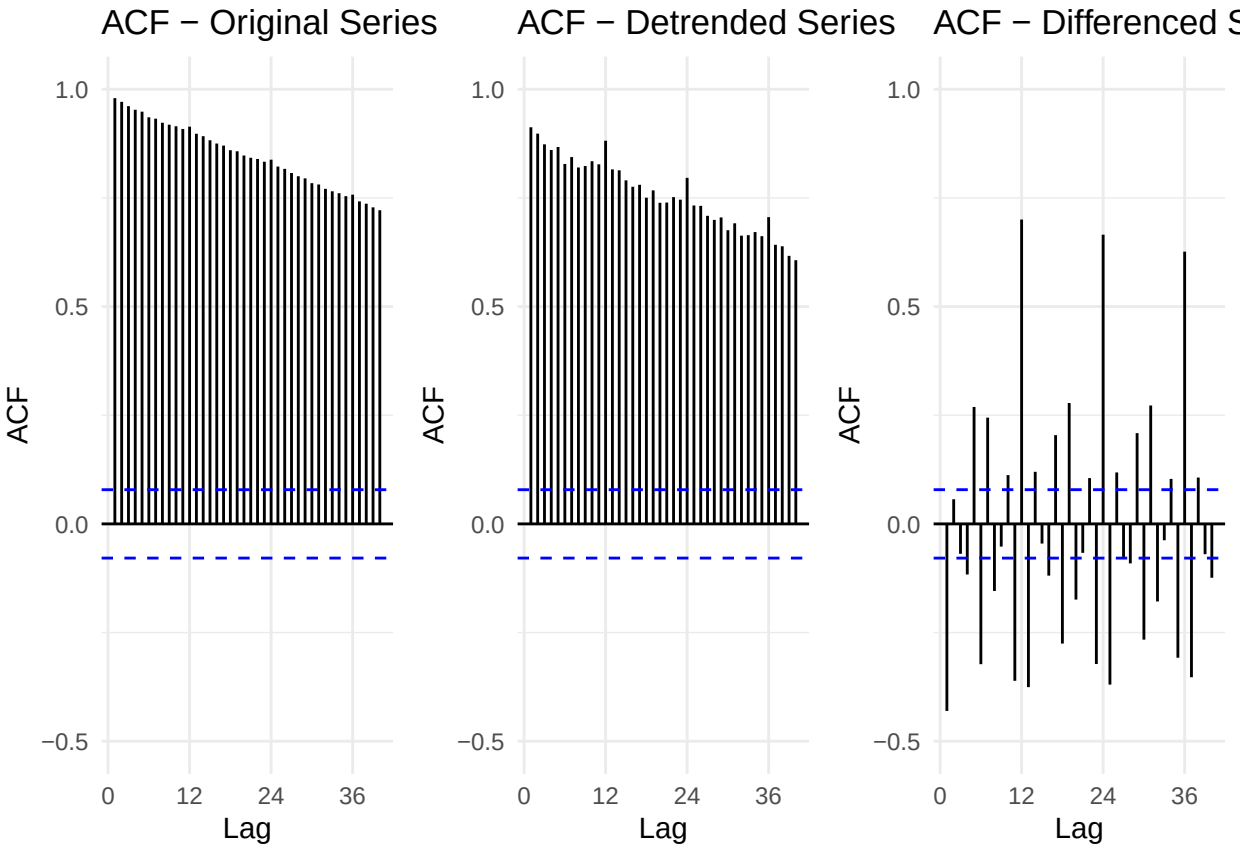
```
# Assign ACF
acf_original <- Acf(ts_data, lag.max = 40, plot = FALSE)
acf_detrend <- Acf(Renewable_detrend_data, lag.max = 40, plot = FALSE)
acf_diff <- Acf(ts_diff, lag.max = 40, plot = FALSE)

# Plot ACF
acf_original_plot <- autoplot(acf_original) +
  ggtitle("ACF - Original Series") +
  ylim(c(-0.5,1)) +
  theme_minimal()

acf_detrend_plot <- autoplot(acf_detrend) +
  ggtitle("ACF - Detrended Series") +
  ylim(c(-0.5,1)) +
  theme_minimal()
```

```r
acf_diff_plot <- autoplot(acf_diff) +
  ggtitle("ACF - Differenced Series") +
  ylim(c(-0.5,1)) +
  theme_minimal()

# Arrange the plots side by side for comparison
plot_grid(acf_original_plot, acf_detrend_plot, acf_diff_plot, ncol = 3)
```




Answer: Differencing is more effective in eliminating the trend. The ACF for linear regrassion still exhibits a slow decay, indicating that the trend was not fully eliminated. By contrast, the ACF for differenced series shows weak autocorrelation at most lags, meaning it is closer to stationarity.

**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
# Perform Seasonal Mann-Kendall Test
summary(SeasonalMannKendall(ts_data))
```

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

```r
# Perform Augmented Dickey-Fuller Test
adf_test <- adf.test(ts_data)
print(adf_test)
```

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

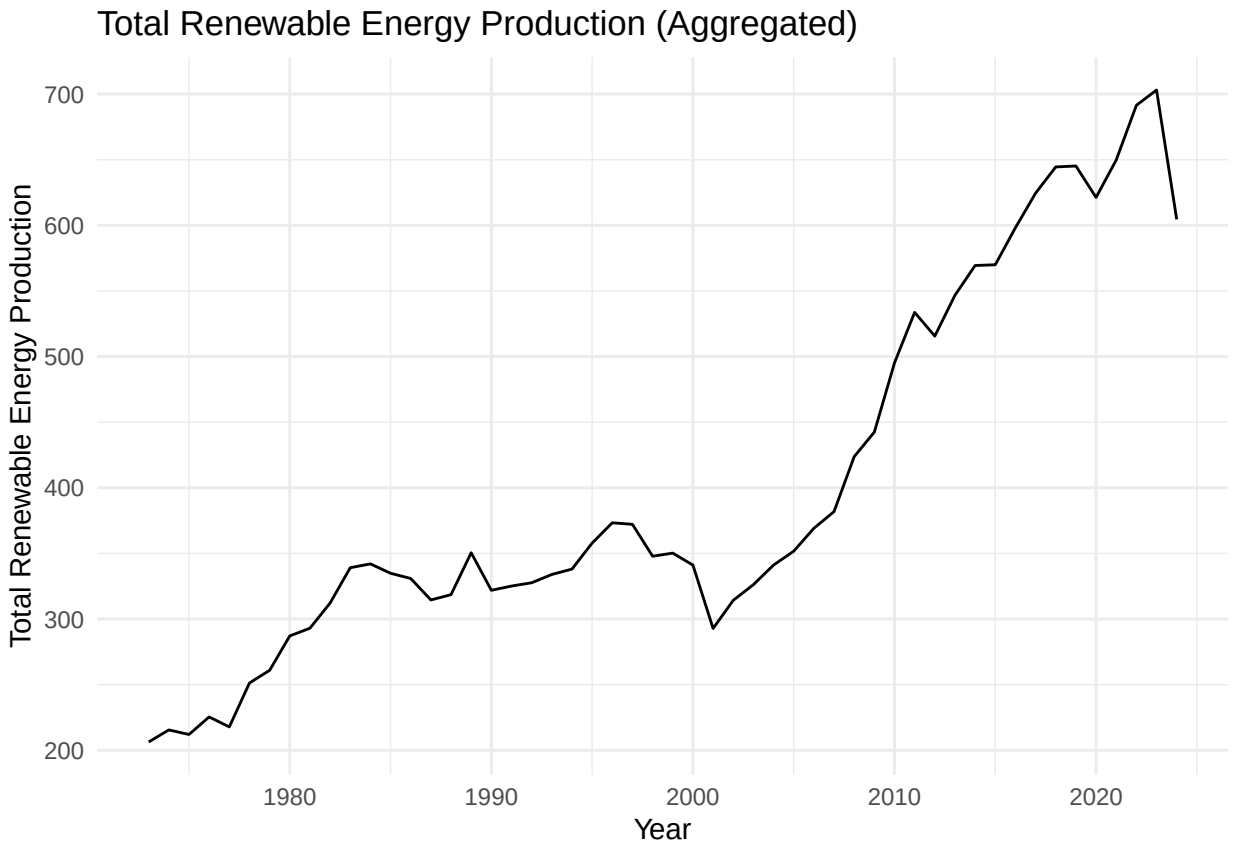
> Answer: From the Seasonal Mann Kendall test, we can reject the null hypothesis and conclude that the series has a significant (upward) trend based on the low p-value which is lower than 0.05. From the ADF test, we fail to reject the null hypothesis with a high p-value (0.9242) – which states that the series has a unit root – and conclude that the series is non stationary and has a stochastic trend. The results do match with what is observed in the Q3 original plot, so differencing helped with removing 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().

```r
# Convert time series into a matrix where rows = months, columns = years
ts_matrix <- matrix(ts_data, nrow = 12, byrow = FALSE)
# Compute annual means (average over 12 months)
annual_means <- colMeans(ts_matrix, na.rm = TRUE)
# Convert to a time series object
ts_annual <- ts(annual_means, start = 1973, frequency = 1)

# Plot the annual aggregated time series
autoplot(ts_annual) +
  ggtitle("Total Renewable Energy Production (Aggregated)") +
  xlab("Year") +
  ylab("Total 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 Q6?

```
summary(MannKendall(ts_annual))
```

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

```
my_year <- c(1973:2024)
cor.test(ts_annual,my_year,method="spearman")
```

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

```
adf.test(ts_annual,alternative = "stationary")
```

```
##
##  Augmented Dickey-Fuller Test
##
```

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

Answer: The p-values from Mann Kendall and Spearman Correlation tests (2.22e-16) are both lower than 0.05, which indicates that we can reject the null hypothesis of stationarity and conclude that the series follows a trend. The p-value from ADF test (0.7098) is greater than 0.05, which indicates that the series contains a unit root and is not stationary. These results for annual series align with the test results for the monthly series.