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

Assignment 4 - Due date 02/10/26

Nicole Gutkowski

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
```

```
#Load/install required package here
library(forecast)
```

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

```
library(tseries)
library(Kendall)
library(cowplot)
library(ggplot2)
library(readxl)
library(openxlsx)
library(lubridate)
```

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

```
## The following object is masked from 'package:cowplot':
##
##     stamp

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

```
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
```

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

energy_data1 =
  read_xlsx(
    path ="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
    skip = 12, sheet="Monthly Data",col_names=FALSE)

#Now let's extract the column names from row 11
read_col_names <-
  read_excel(
    path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
    skip = 10,n_max = 1, sheet="Monthly Data",col_names=FALSE)

#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
#subset for the cols that we want
subset_df = energy_data1[,c("Month",
                            "Total Renewable Energy Production")]
subset_df$Month <- as.Date(subset_df$Month)

head(subset_df)
```

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

```r
#clean so there are an even number of months/ years
nobs <- nrow(subset_df)

cleaned_df <- subset_df[1:(nobs-9),]


#Tail again to check if the rows were correctly removed
tail(cleaned_df)
```

```
## # A tibble: 6 x 2
##   Month      `Total Renewable Energy Production`
##   <date>                                   <dbl>
## 1 2024-07-01                                757.
## 2 2024-08-01                                756.
## 3 2024-09-01                                700.
## 4 2024-10-01                                735.
## 5 2024-11-01                                726.
## 6 2024-12-01                                742.
```

Stochastic Trend and Stationarity Tests

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

Q1

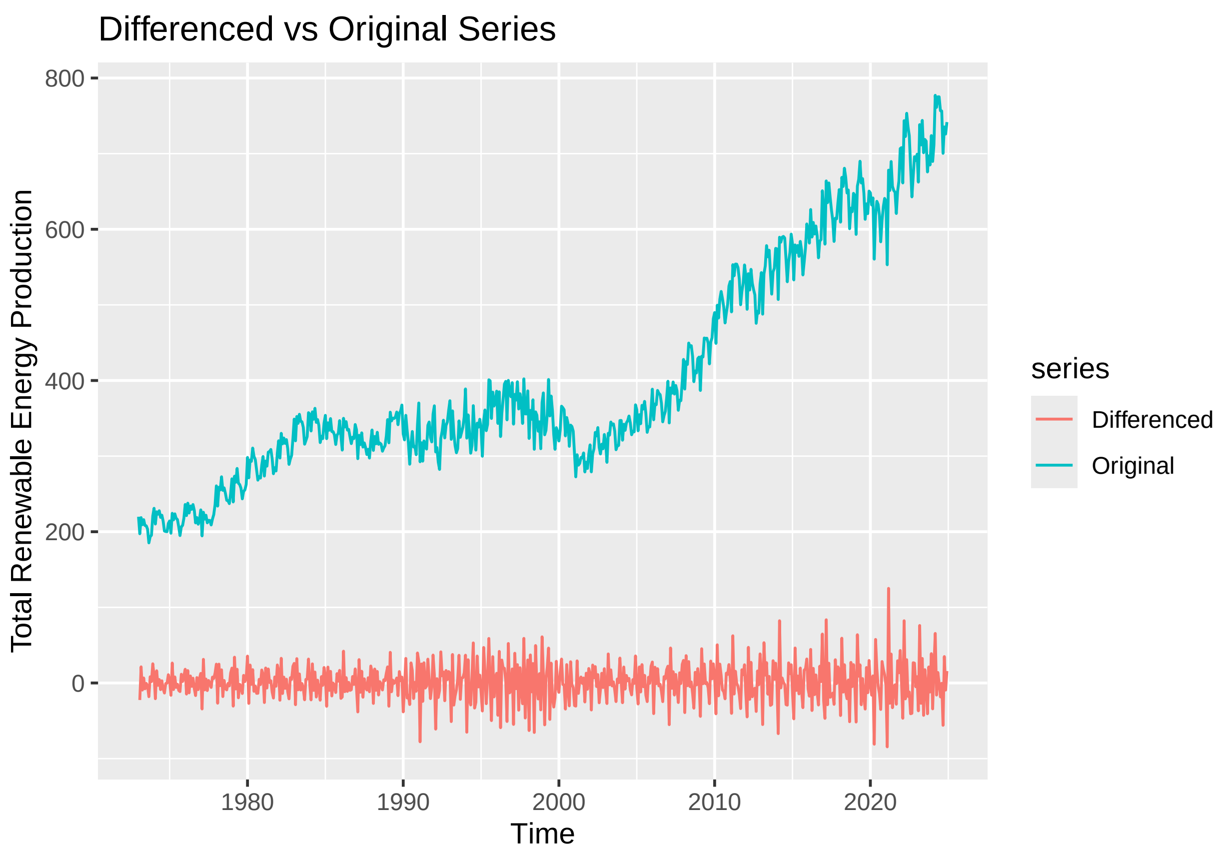
Difference the "Total Renewable Energy Production" series using function diff(). Function diff() is from package base and take three main arguments: * x vector containing values to be differenced; * *lag* integer indicating with lag to use; * *differences* integer indicating how many times series should 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_data = ts(cleaned_df[,2],frequency=12, start = c(1973,1))

ts_diff = diff(ts_data, lag = 1, differences = 1)

autoplot(ts_diff, series = "Differenced")+
  autolayer(ts_data, series = "Original")+
  labs(title = "Differenced vs Original Series",
       y = "Total Renewable Energy Production")
```



Answer: After differencing the series, there no longer appears to be a trend.

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.

```r
nobs <- nrow(cleaned_df)
t <- c(1:nobs)

energy_data = as.data.frame(cleaned_df)



#linear trend for Total Renew
linear_reg <- lm(cleaned_df[[2]]~t)
#regressing over the time vector just created
summary(linear_reg)
```

```
##
## Call:
## lm(formula = cleaned_df[[2]] ~ t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -152.38  -38.35   12.95   41.57  151.49
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 175.08397    5.02585   34.84   <2e-16 ***
## t             0.73268    0.01393   52.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.7 on 622 degrees of freedom
## Multiple R-squared:  0.8164, Adjusted R-squared:  0.8161
## F-statistic:  2765 on 1 and 622 DF,  p-value: < 2.2e-16
```

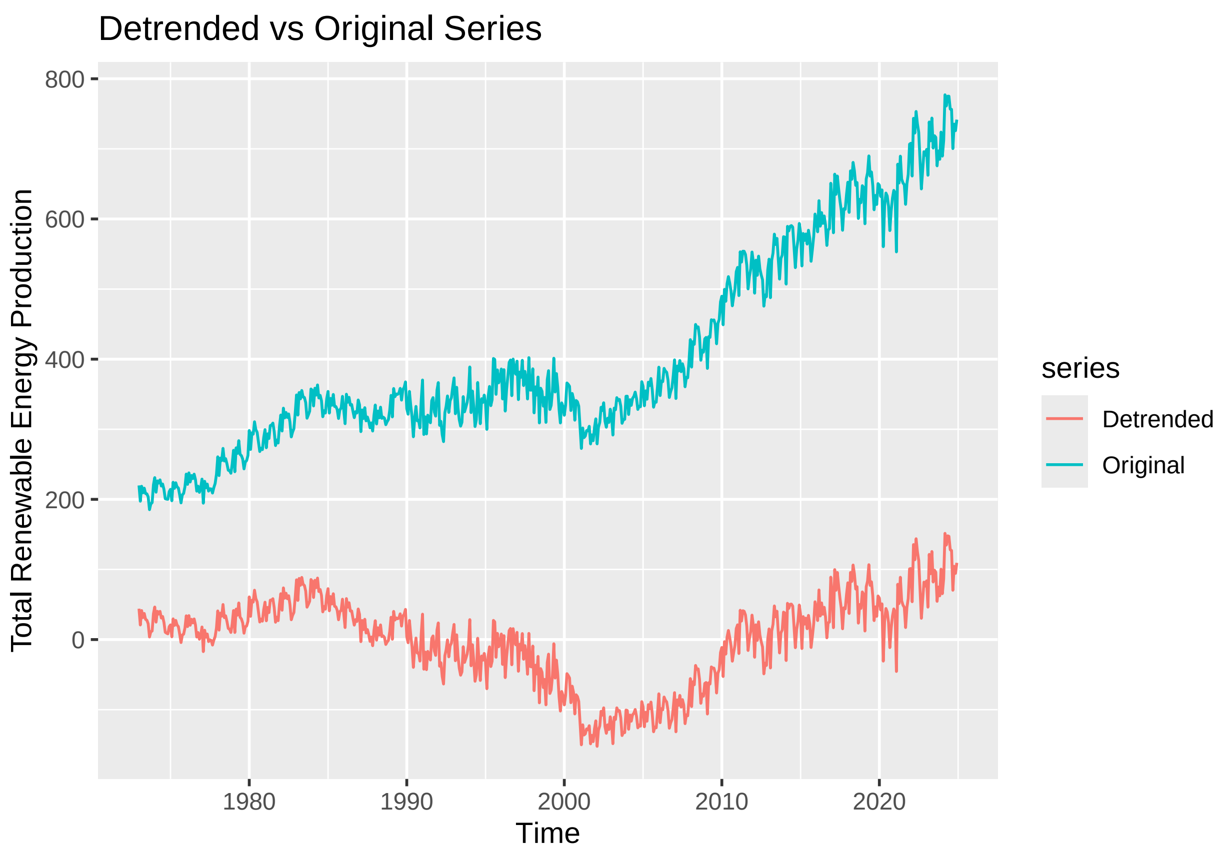
```r
#store the reg coefficients
beta1 <- as.numeric(linear_reg$coefficients[1]) #intercept
beta2 <- as.numeric(linear_reg$coefficients[2]) #slope

#Total Renewable
#remove the trend from series total series
detrend_data <- cleaned_df[[2]] - (beta1 + beta2*t)
#subtracting the equation (with the stored coeffs and the t value)
class(detrend_data)
```

```
## [1] "numeric"
```

```r
ts_detrend <- ts(detrend_data, frequency = 12, start = c(1973,1))


autoplot(ts_data, series = "Original")+
  autolayer(ts_detrend, series = "Detrended")+
  labs(title = "Detrended vs Original Series",
    y = "Total Renewable Energy Production")
```



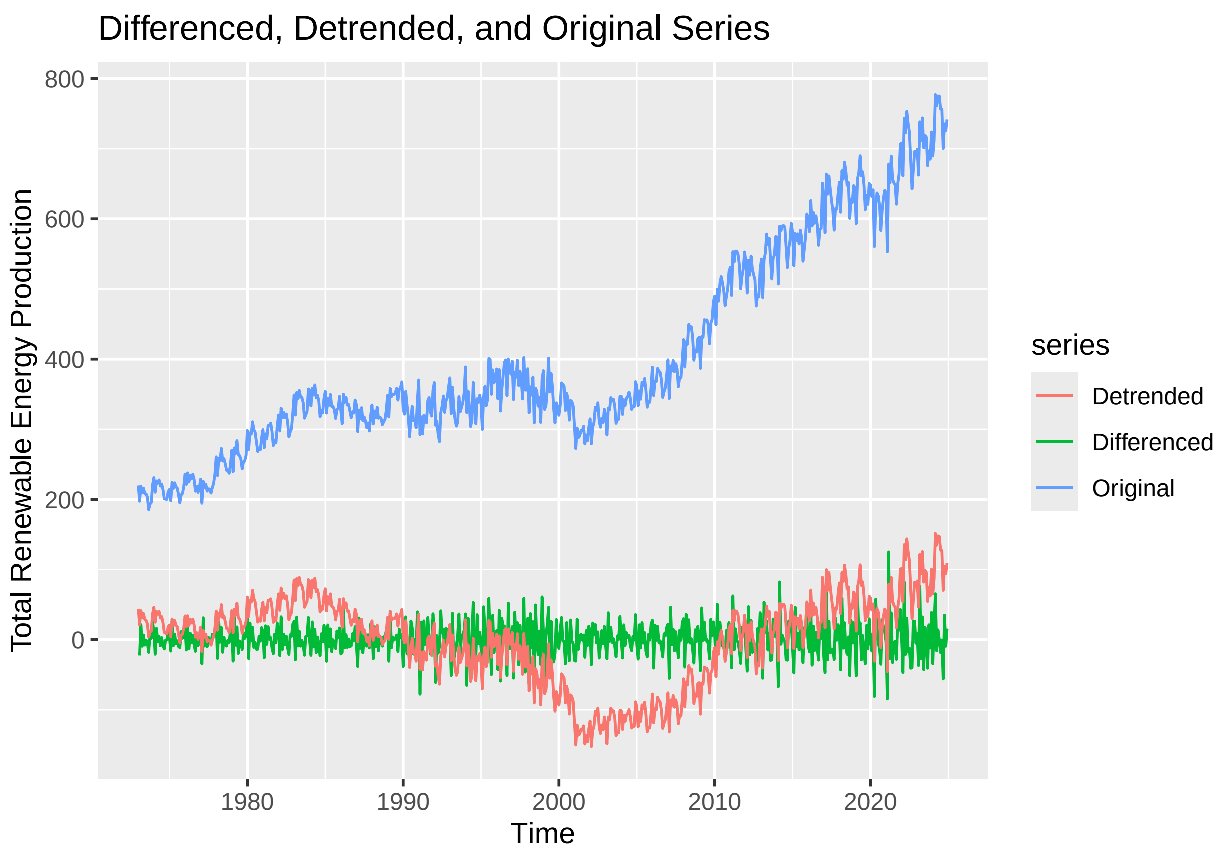
Q3

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?

```
autoplot(ts_diff, series = "Differenced")+
  autolayer(ts_data, series = "Original")+
  autolayer(ts_detrend, series = "Detrended")+
  labs(title = "Differenced, Detrended, and Original Series",
       y = "Total Renewable Energy Production")
```

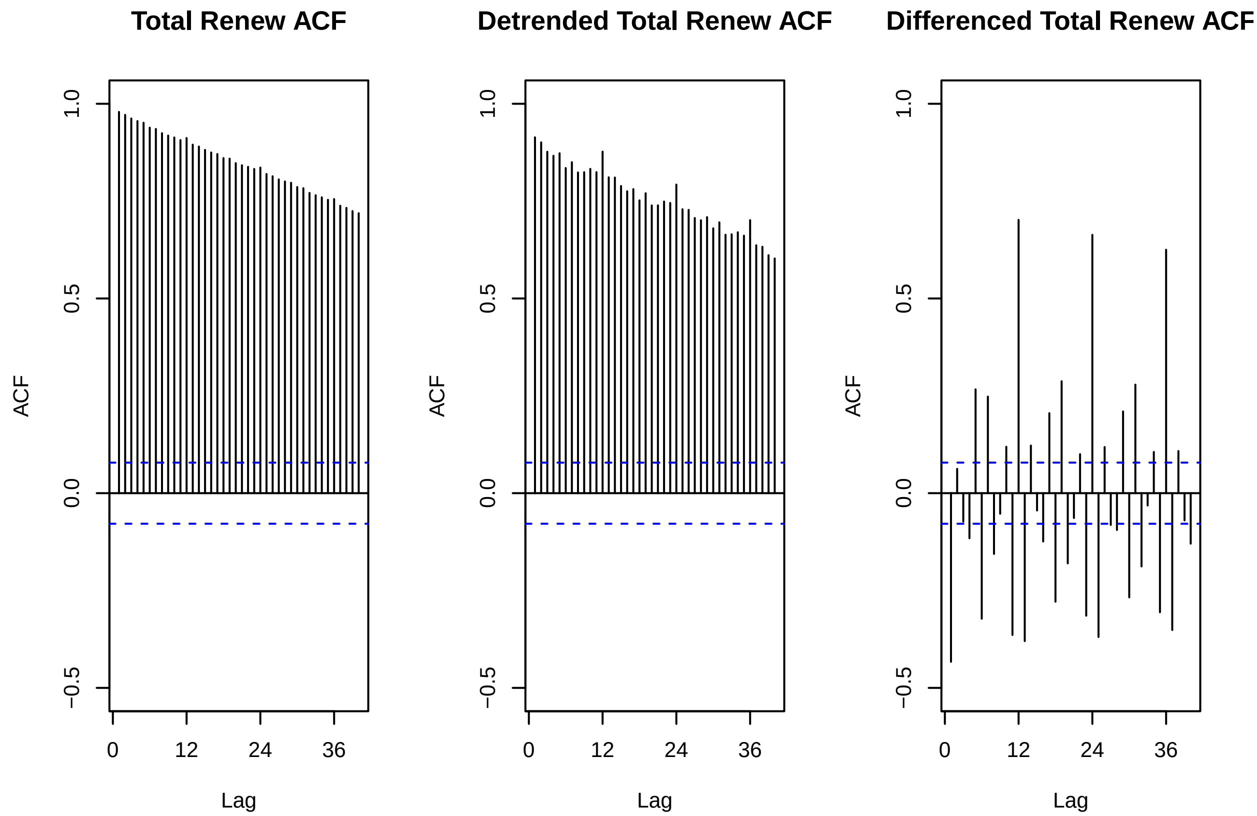


Answer: The differencing method seems to have been more efficient in removing the trend. In the detrended plot, there is still a general up and down behavior in the plot, however the differenced series lies virtually flat on the 0 of the y-axis.

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?

```
#ACF Comparison
par(mfrow=c(1,3))  #place plot side by side
Acf(ts_data,lag.max=40,main=paste("Total Renew ACF",sep=""), ylim=c(-0.5,1))
Acf(ts_detrend,lag.max=40,main=paste("Detrended Total Renew ACF",sep=""), ylim=c(-0.5,1))
Acf(ts_diff,lag.max=40,main=paste("Differenced Total Renew ACF",sep=""), ylim=c(-0.5,1))
```



Answer: The differencing method was more efficient in eliminating the trend for the series. While detrending the series did slightly drop the ACF values and reveal possible seasonal patterns through minor spikes at the 12, 24, and 36 lag distances, the differenced series has much lower ACF values almost immediately and the potential seasonal pattern is indicated through clear spikes at the 12, 24, and 36 lag distances.

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

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

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

```
## Score =  12617 , Var(Score) = 192711
## denominator =  15911.5
```

```
## tau = 0.793, 2-sided pvalue =< 2.22e-16
## NULL
```

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

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

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

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_data
## Dickey-Fuller = -1.0895, Lag order = 8, p-value = 0.9243
## alternative hypothesis: stationary
```

> Answer: The seasonal Mann Kendall test shows a significant upwards trend with a tau value of 0.793 and a p-value less than 0.05. From the ADF test, we fail to reject that null hypothesis that the series has a unit root, indicating that the series is non-stationary and follows a stochastic 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().

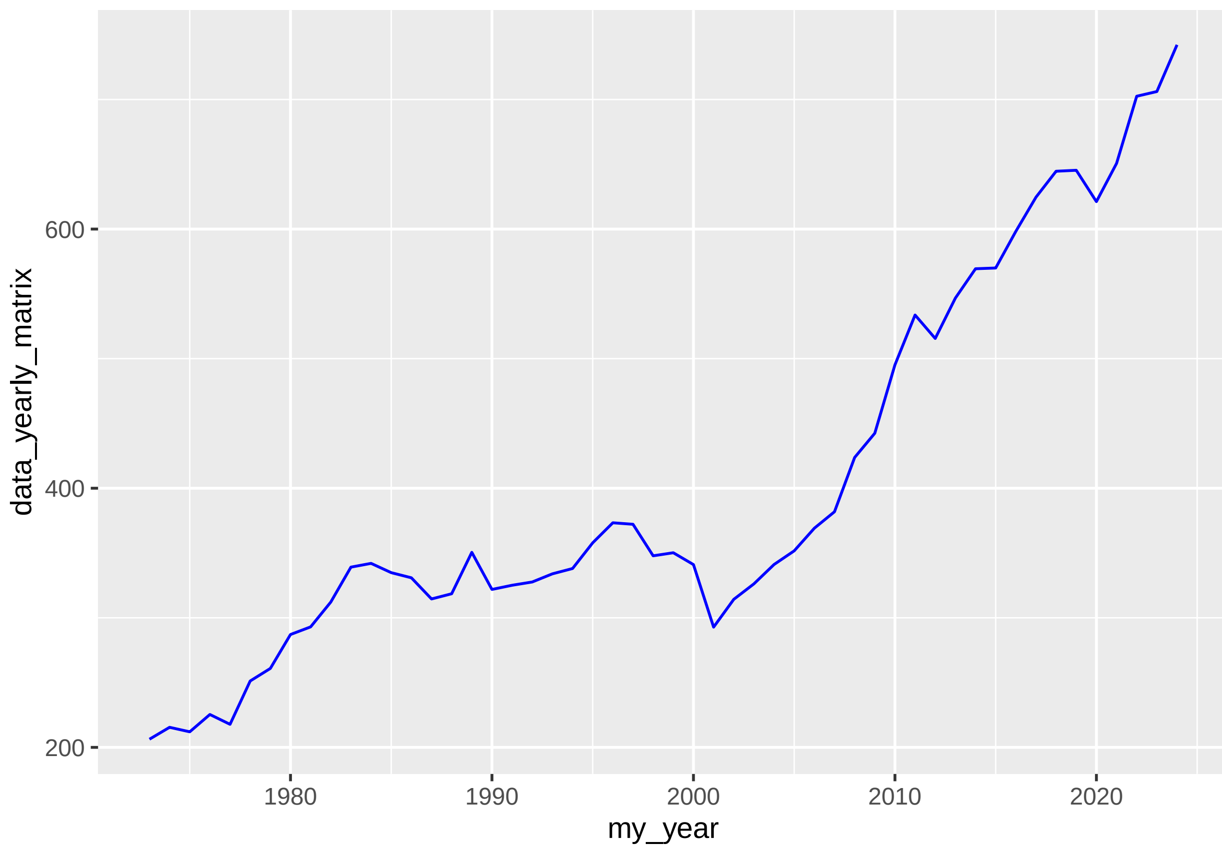
```
#Group data in yearly steps instances
data_matrix <- matrix(ts_data,byrow=FALSE,nrow=12)
#each year is a column
data_yearly_matrix <- colMeans(data_matrix)
#calculating the average by year (matrix columns)

## if you want to run the mann kendall test --> need to remove seasonality
## converting to a matrix

#creating a year series
my_year <- c(year(first(cleaned_df$Month)):year(last(cleaned_df$Month)))

#combining the years + data series into a df
new_year_data <- data.frame(my_year, data_yearly_matrix)

#plotting to check results
ggplot(new_year_data, aes(x=my_year, y=data_yearly_matrix)) +
            geom_line(color="blue")
```



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?

```
#Use yearly date to run Mann Kendall
print("Results of Mann Kendall on average yearly series")
```

```
## [1] "Results of Mann Kendall on average yearly series"
```

```
print(summary(MannKendall(data_yearly_matrix)))
```

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

```
#Spearman Correlation Rank Test
sp_rho=cor.test(new_year_data$data_yearly_matrix,
  new_year_data$my_year, method="spearman")
print(sp_rho)
```

```
##
```

```
##  Spearman's rank correlation rho
##
## data:  new_year_data$data_yearly_matrix and new_year_data$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
#Null hypothesis is that data has a unit root
print("Results for ADF test/n")
```

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

```r
print(adf.test(new_year_data$data_yearly_matrix,alternative = "stationary"))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  new_year_data$data_yearly_matrix
## Dickey-Fuller = -0.85301, Lag order = 3, p-value = 0.9515
## alternative hypothesis: stationary
```

> Answer: Both the Mann Kendall and Spearman tests indicate that there is a significant strong increasing trend in the series. Both tests have p-values of less than 0.05 and a tau and rho of 0.817 and 0.921 respectively. The ADF test p value of 0.95 indicates that we cannot reject the null-hypothesis that the series is non-stationary. These results suggest that the series has a strong increasing, non-stationary trend. These findings are consistent with the results in Q5.