ENV 790.30 - Time Series Analysis for Energy Data | Spring 2024

Assignment 4 - Due date 02/12/24

David Robinson

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(forecast)
```

```
## Warning: package 'forecast' was built under R version 4.3.2

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

```
library(tseries)
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
```

```r
library(readxl)
library(ggplot2)
library(Kendall)
library(cowplot)
```

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".

```r
#Importing data set - using readxl package

getwd()
```

```
## [1] "C:/Users/dhr20/OneDrive - Duke University/1 - Academics/1 - First Year/2 - Spring 2024/3 - Time
```

```r
raw_energy_data <- read_excel(path="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sou
```

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

```r
colnames(raw_energy_data)=c("Month",
                             "Wood Energy Production",
                             "Biofuels Production",
                             "Total Biomass Energy Production",
                             "Total Renewable Energy Production",
                             "Hydroelectric Power Consumption",
                             "Geothermal Energy Consumption",
                             "Solar Energy Consumption",
                             "Wind Energy Consumption",
                             "Wood Energy Consumption",
                             "Waste Energy Consumption",
                             "Biofuels Consumption",
                             "Total Biomass Energy Consumption",
                             "Total Renewable Energy Consumption")
```

```r
raw_energy_data <- raw_energy_data[,1:6]
raw_energy_data_dates <- raw_energy_data[,1]
raw_energy_data_renewable <- raw_energy_data[,5]
raw_energy_data <- cbind(raw_energy_data_dates,raw_energy_data_renewable)

head(raw_energy_data)
```

```
##        Month Total Renewable Energy Production
## 1 1973-01-01                           219.839
## 2 1973-02-01                           197.330
## 3 1973-03-01                           218.686
## 4 1973-04-01                           209.330
## 5 1973-05-01                           215.982
## 6 1973-06-01                           208.249
```

```r
nobs <- nrow(raw_energy_data)

t <- 1:nobs

ts_renewable <- ts(raw_energy_data[t,1:2], frequency=12,start=c(1973,1))

head(ts_renewable)
```

```
##              Month Total Renewable Energy Production
## Jan 1973  94694400                           219.839
## Feb 1973  97372800                           197.330
## Mar 1973  99792000                           218.686
## Apr 1973 102470400                           209.330
## May 1973 105062400                           215.982
## Jun 1973 107740800                           208.249
```

Stochastic Trend and Stationarity Tests

Q1

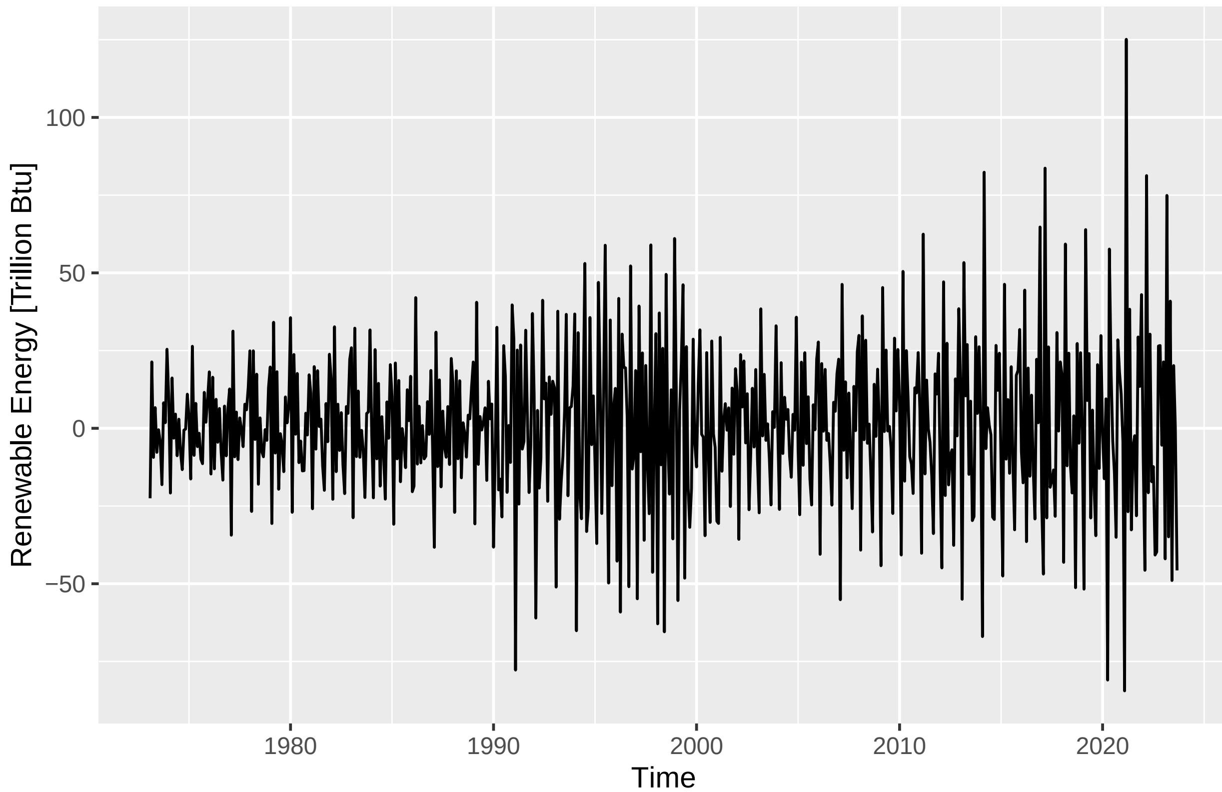
Difference the "Total Renewable Energy Production" series using function diff(). Function diff() is from package base and take three main arguments: * *x* vector containing values to be differenced; * *lag* integer indicating with lag to use; * *differences* integer indicating how many times series should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

ts_renewable_diff_plot <- autoplot(ts_renewable_diff) +
                                    ylab("Renewable Energy [Trillion Btu]")+
                                    ggtitle("")

ts_renewable_diff_plot
```




```
#The series does not seem to still have 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 the same name for you time series object that you had in A3.

```
renewable_linear_trend <- lm(raw_energy_data[,2]~t)
summary(renewable_linear_trend)
```

```
##
## Call:
## lm(formula = raw_energy_data[, 2] ~ t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -148.27  -35.63   11.58   41.51  144.27
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 180.98940    4.90151   36.92   <2e-16 ***
## t             0.70404    0.01392   50.57   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.41 on 607 degrees of freedom
## Multiple R-squared:  0.8081, Adjusted R-squared:  0.8078
## F-statistic:  2557 on 1 and 607 DF,  p-value: < 2.2e-16
```

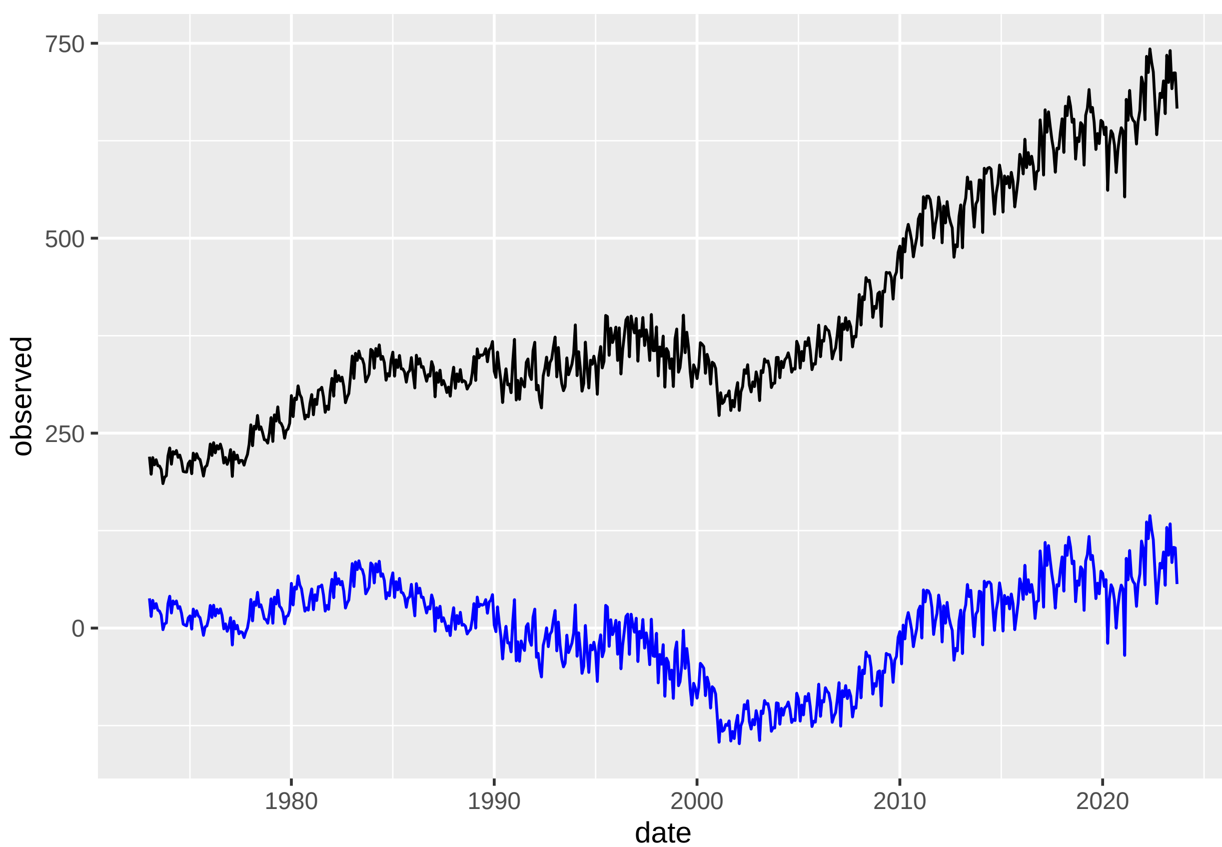
```
renewable_beta0 <- renewable_linear_trend$coefficients[1]
renewable_beta1 <- renewable_linear_trend$coefficients[2]

renewable_y_detrend <- raw_energy_data[,2] -
  (renewable_beta0 + renewable_beta1*t)

renewable_df_detrend <- data.frame("date"=raw_energy_data[,1],
                                   "observed"=raw_energy_data[,2],
                                   "detrend"=renewable_y_detrend)

ts_renewable_detrend <- ts(renewable_y_detrend, frequency=12,start=c(1973,1))

ggplot(renewable_df_detrend,aes(x=date))+
  geom_line(aes(y=observed),color="black")+
  geom_line(aes(y=detrend),color="blue")
```

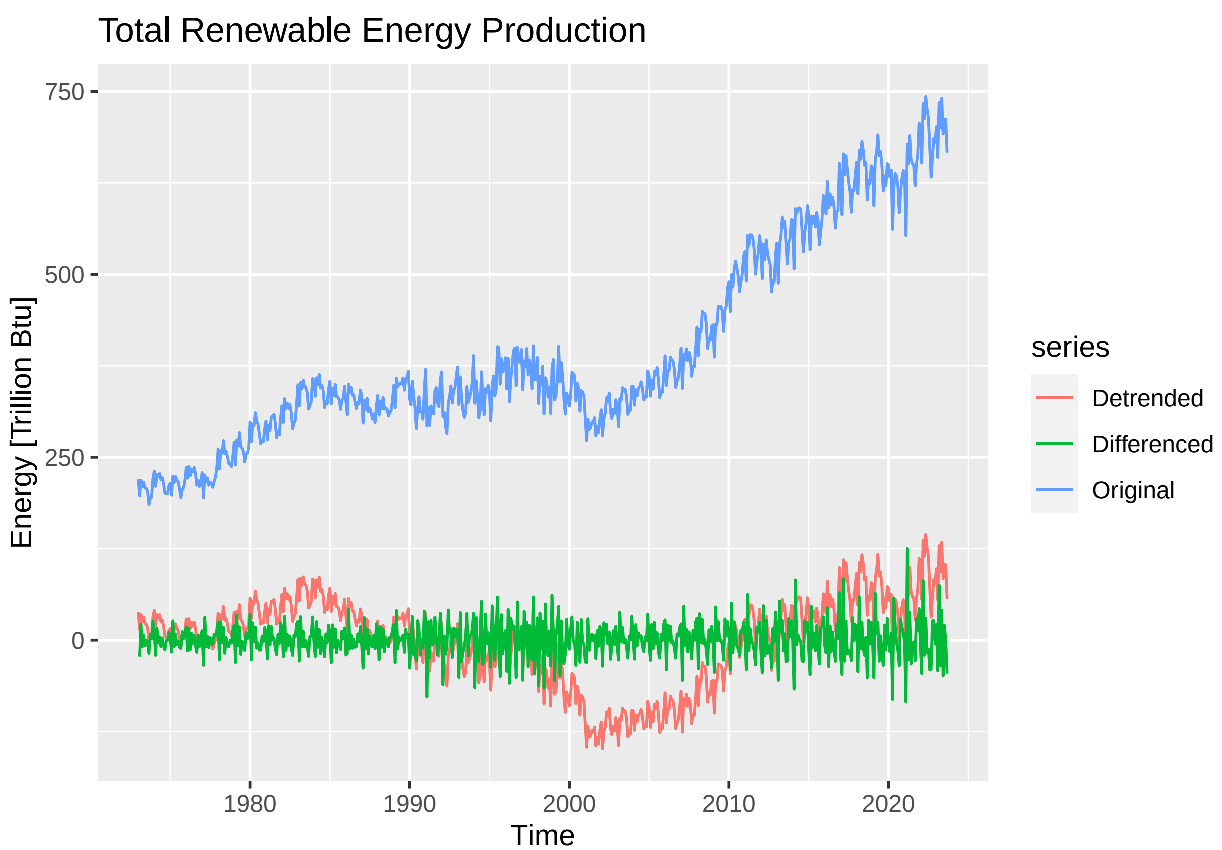


Q3

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.

```
autoplot(ts_renewable[,2], series="Original")+
  autolayer(ts_renewable_detrend, series="Detrended")+
  autolayer(ts_renewable_diff, series="Differenced")+
  ylab("Energy [Trillion Btu]")+
  ggtitle("Total Renewable Energy Production")
```



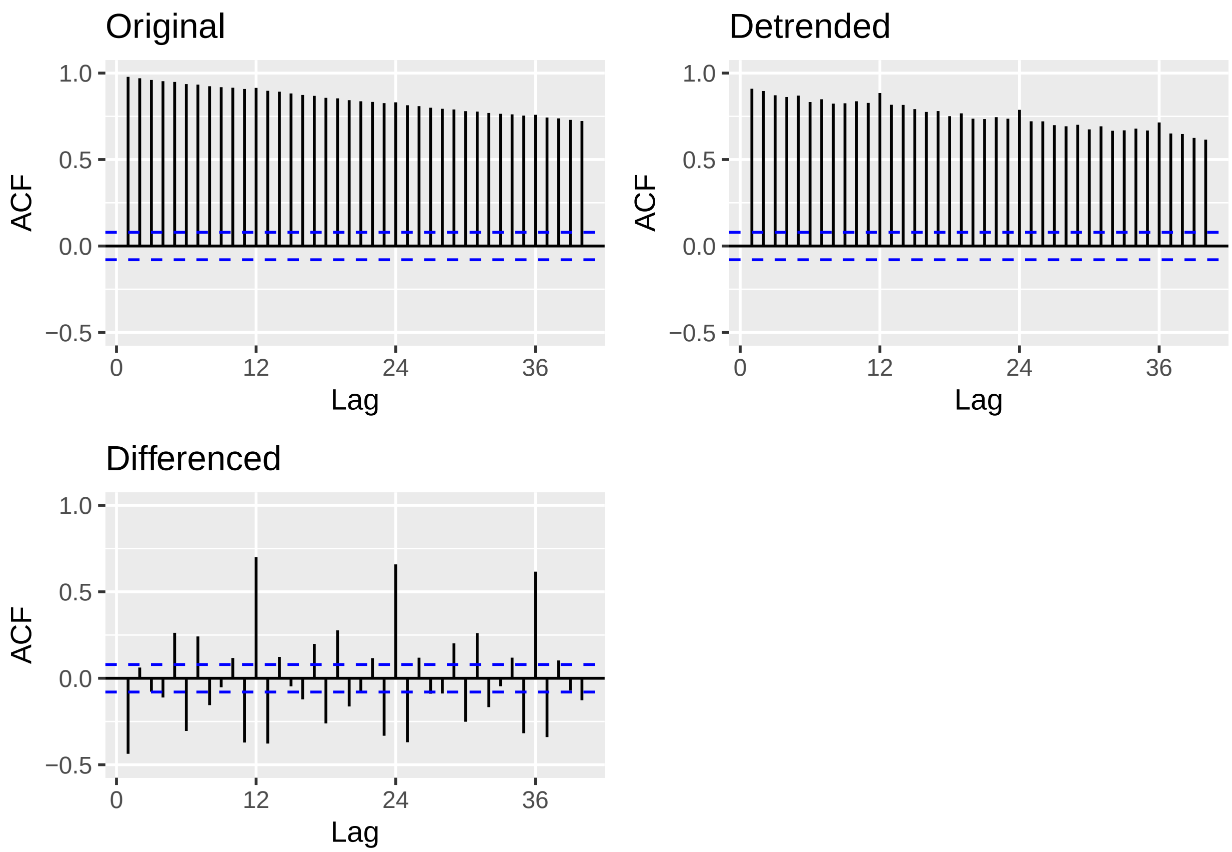
Q4

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. Which method do you think was more efficient in eliminating the trend? The linear regression or differencing?

```
plot_grid(
  autoplot(Acf(ts_renewable[,2], lag.max=40,plot=FALSE),main="Original",
```

```r
            ylim=c(-0.5,1)),
  autoplot(Acf(ts_renewable_detrend, lag.max=40,plot=FALSE),main="Detrended",
            ylim=c(-0.5,1)),
  autoplot(Acf(ts_renewable_diff,lag.max=40,plot=FALSE),main="Differenced",
            ylim=c(-0.5,1))
)
```

```
## 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‘
```




```r
#The differencing appears to have been more effective in eliminating the trend
#-- the correlation values are lower for the Differenced ACF than for the
#Detrended ACF. The Differenced ACF indicates some seasonality given spikes at
#lags 12, 24, and 36.
```

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 Q1? 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.

```r
SMKtest_original <- SeasonalMannKendall(ts_renewable[,2])
print("Results for Seasonal Mann Kendall /n")
```

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

```r
print(summary(SMKtest_original))
```

```
## Score =  11865 , Var(Score) = 179299
## denominator =  15149.5
## tau = 0.783, 2-sided pvalue =< 2.22e-16
## NULL
```

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

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

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

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_renewable[, 2]
## Dickey-Fuller = -1.24, Lag order = 8, p-value = 0.9
## alternative hypothesis: stationary
```

```r
#The conclusion for the Seasonal Mann Kendall test is that there is a
#significant seasonal trend in the data (given the p-value less than 0.05).
#Because the S statistic is positive, it suggests an increasing seasonal trend.
#S is relatively large so the trend is relatively strong.

#The conclusion for the ADF test, given the p-value of 0.9, is that we fail to
#reject the null hypothesis. This suggests that the time series may have a unit
#root and is non-stationary.
```

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 accumulated yearly series into a time series object and plot the series using autoplot().

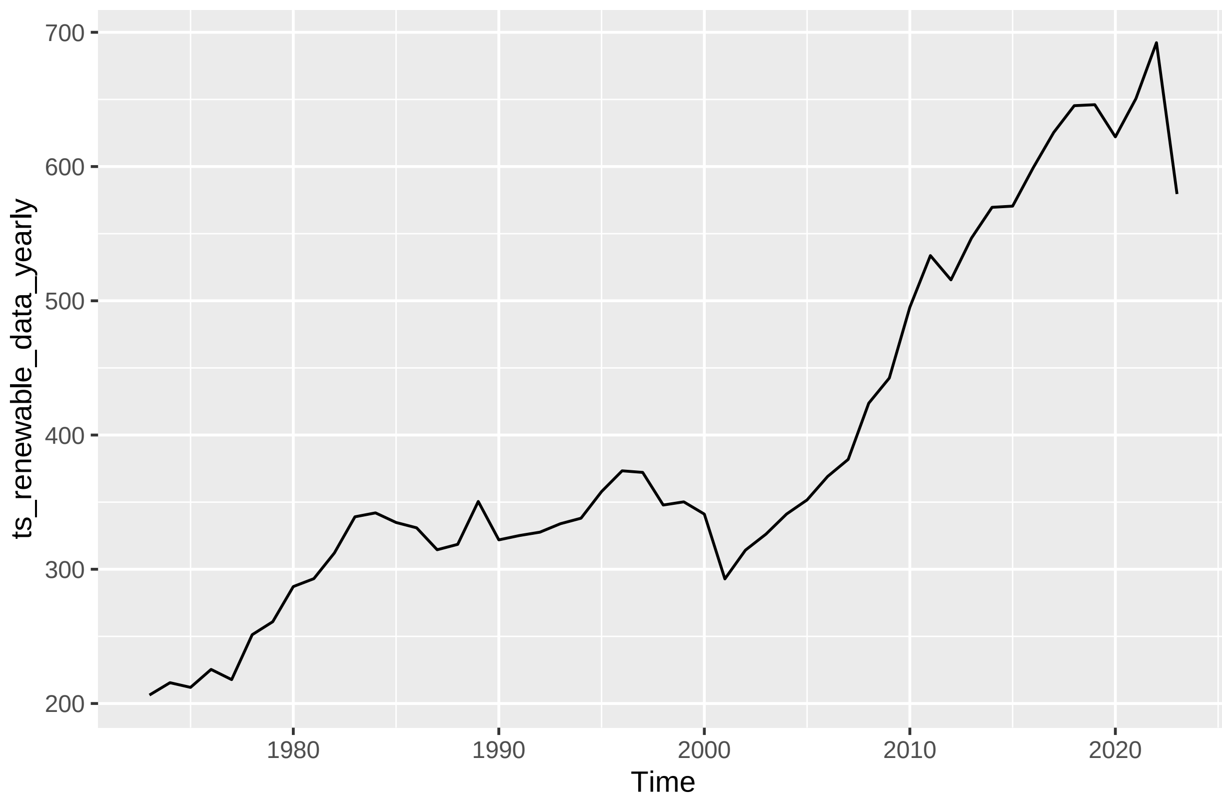
```r
renewable_data_matrix <- matrix(ts_renewable[,2],byrow=FALSE,nrow=12)
```

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

```
renewable_data_yearly <- colMeans(renewable_data_matrix)

ts_renewable_data_yearly <- ts(renewable_data_yearly,start = 1973,
                               frequency = 1 )

autoplot(ts_renewable_data_yearly)
```



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 data 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(ts_renewable_data_yearly)))
```

```
## Score =  1019 , Var(Score) = 15158.33
## denominator =  1275
## tau = 0.799, 2-sided pvalue =< 2.22e-16
## NULL
```

```r
#Deterministic trend with Spearman Correlation Test
print("Results from Spearman Correlation")
```

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

```r
sp_rho=cor(ts_renewable_data_yearly,c(1973:2023),method="spearman")
print(sp_rho)
```

```
## [1] 0.9136652
```

```r
#Now let's try the yearly data
print("Results for ADF test on yearly data/n")
```

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

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

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_renewable_data_yearly
## Dickey-Fuller = -2.0953, Lag order = 3, p-value = 0.5361
## alternative hypothesis: stationary
```

```r
#The results of the Mann-Kendall test are in alignment with those from Q5 --
# low p-value so we reject the null and conclude that there is a trend. S is
#positive so it's an increasing trend.

#We did not run the Spearman correlation test for Q5, so there is no basis
#for comparison. That said, the correlation coefficient value of 0.9137
#suggests a strong positive monotonic relationship between the yearly time
#series data and the sequence of years.

#The conclusion for the ADF test, given the p-value of 0.5, is that we fail to
#reject the null hypothesis. This suggests that the time series may have a unit
#root and is non-stationary. This is similar to Q5.
```