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

Assignment 4 - Due date 02/12/24

Sai Powar

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)
```

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

```
library(tseries)
library(Kendall)
library(lubridate)
```

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

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

```
library(ggplot2)
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
```

```
library(base)
library(cowplot)
```

```
##
## Attaching package: 'cowplot'

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

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
raw_energy_data <- read.table(file="./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Sou

#date
energy_date <- ym(raw_energy_data[,1])  #function my from package lubridate
head(energy_date)
```

```
## [1] "1973-01-01" "1973-02-01" "1973-03-01" "1973-04-01" "1973-05-01"
## [6] "1973-06-01"
```

```
re_data <- cbind(energy_date,raw_energy_data[,5, drop=FALSE])
head(re_data)
```

```
##   energy_date 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
```

```
nobs_re <- nrow(re_data)
t <- c(1:nobs_re)

#creating time series object
ts_re_data <- ts(re_data[,2],start = c(1973,1),frequency=12)
head(ts_re_data)
```

```
##          Jan     Feb     Mar     Apr     May     Jun
## 1973 219.839 197.330 218.686 209.330 215.982 208.249
```

Stochastic Trend and Stationarity Tests

Q1

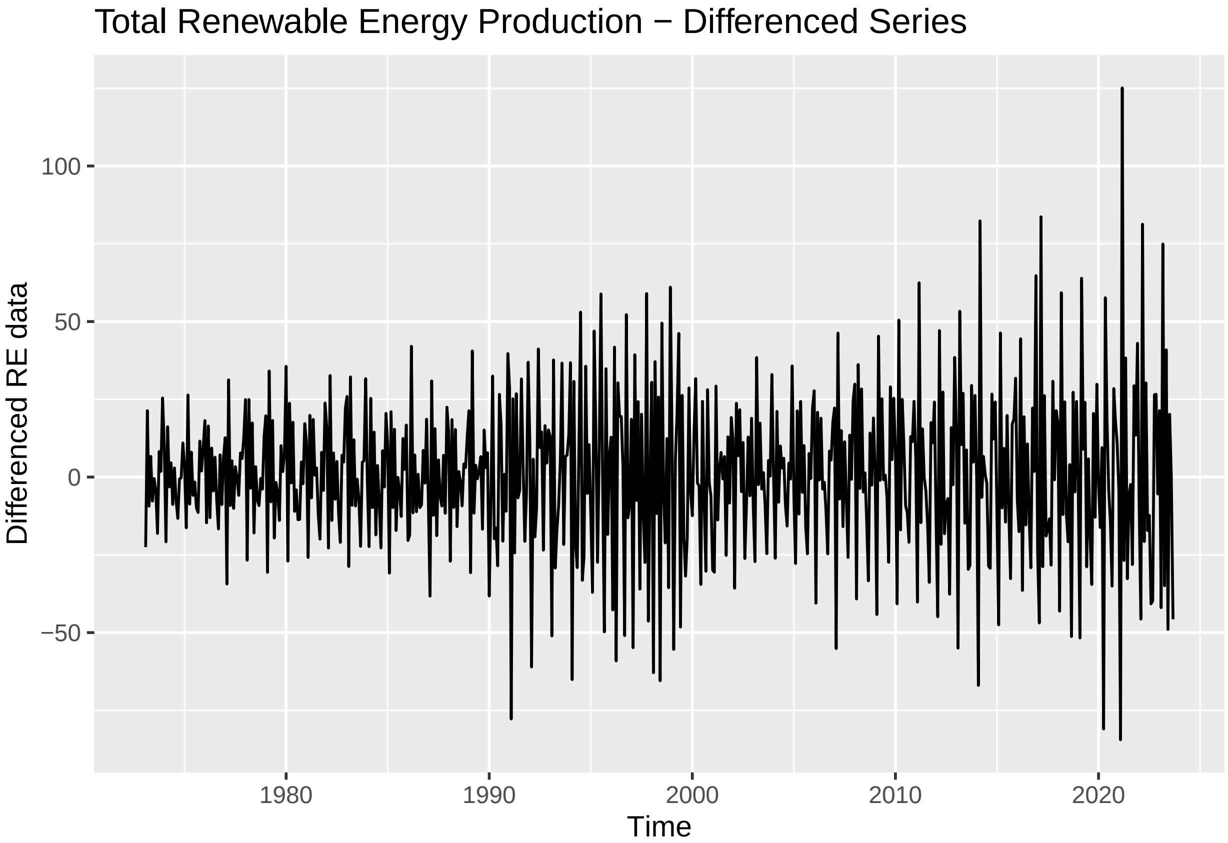
Difference the "Total Renewable Energy Production" series using function diff(). Function diff() is from package base and take three main arguments: * *x* vector containing values to be differenced; * *lag* integer indicating with lag to use; * *differences* integer indicating how many times series should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

```
diff_re_data <- diff(ts_re_data, lag=1, differences = 1)
head(diff_re_data)
```

```
##          Feb     Mar     Apr     May     Jun     Jul
## 1973 -22.509  21.356  -9.356   6.652  -7.733  -0.449
```

```
autoplot(diff_re_data,ylab="Differenced RE data", xlab = "Time", main = "Total Renewable Energy Producti
```



The increasing linear trend present in the original series is not seen in the differenced series.

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 orinal 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.

```
re_linear = lm(ts_re_data~t, cbind(ts_re_data,t))
summary(re_linear)
```

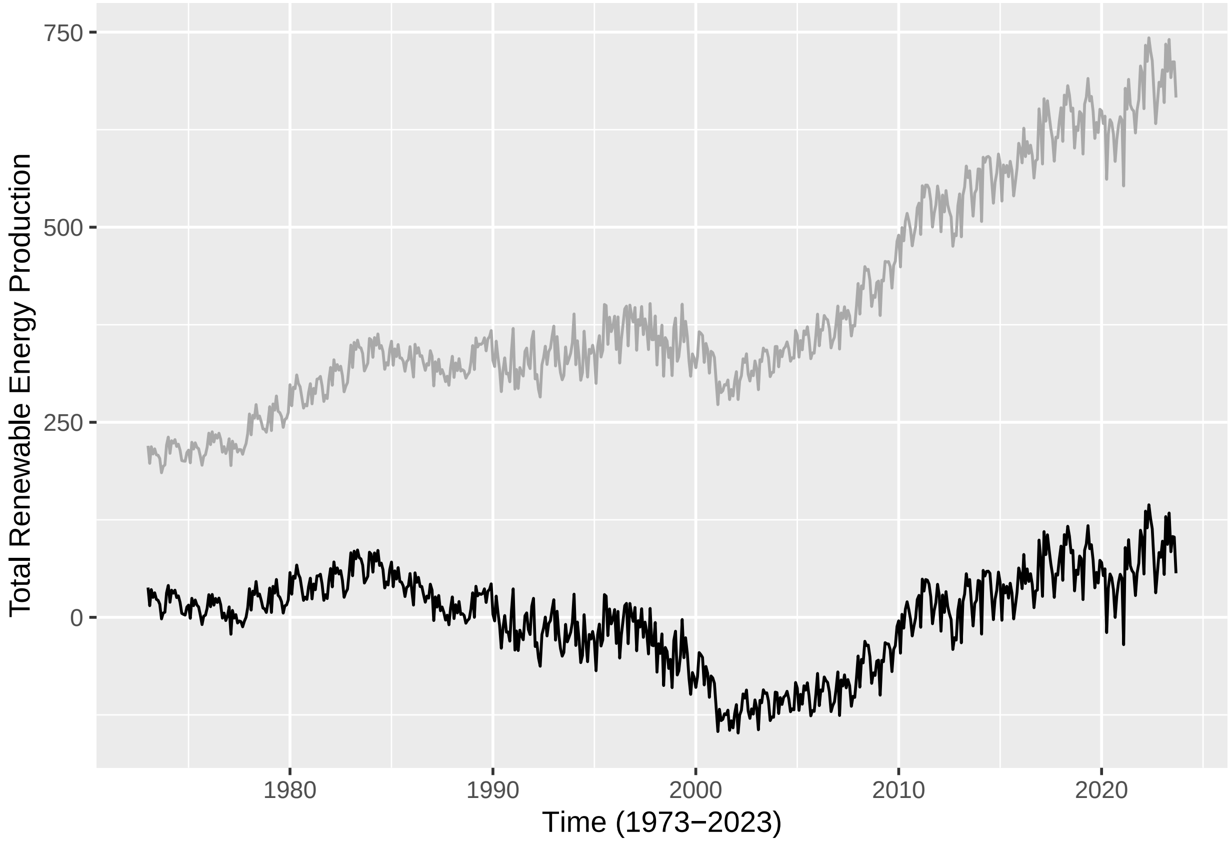
```
##
## Call:
## lm(formula = ts_re_data ~ t, data = cbind(ts_re_data, t))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -148.27  -35.63   11.58   41.51  144.27
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 180.98940    4.90151   36.92   <2e-16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

```
re_beta0=as.numeric(re_linear$coefficients[1])
re_beta1=as.numeric(re_linear$coefficients[2])
print(summary(re_linear))
```

```
##
## Call:
## lm(formula = ts_re_data ~ t, data = cbind(ts_re_data, 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
```

```
dt_re_linear <- ts_re_data-(re_beta0+re_beta1*t)

p2 <- ggplot(re_data, aes(x=energy_date, y=re_data[,2])) +
          geom_line(color="darkgrey") +
          ylab("Total Renewable Energy Production") +
          xlab("Time (1973-2023)")+
          geom_line(aes(y=dt_re_linear), col="black")
p2
```



Q3

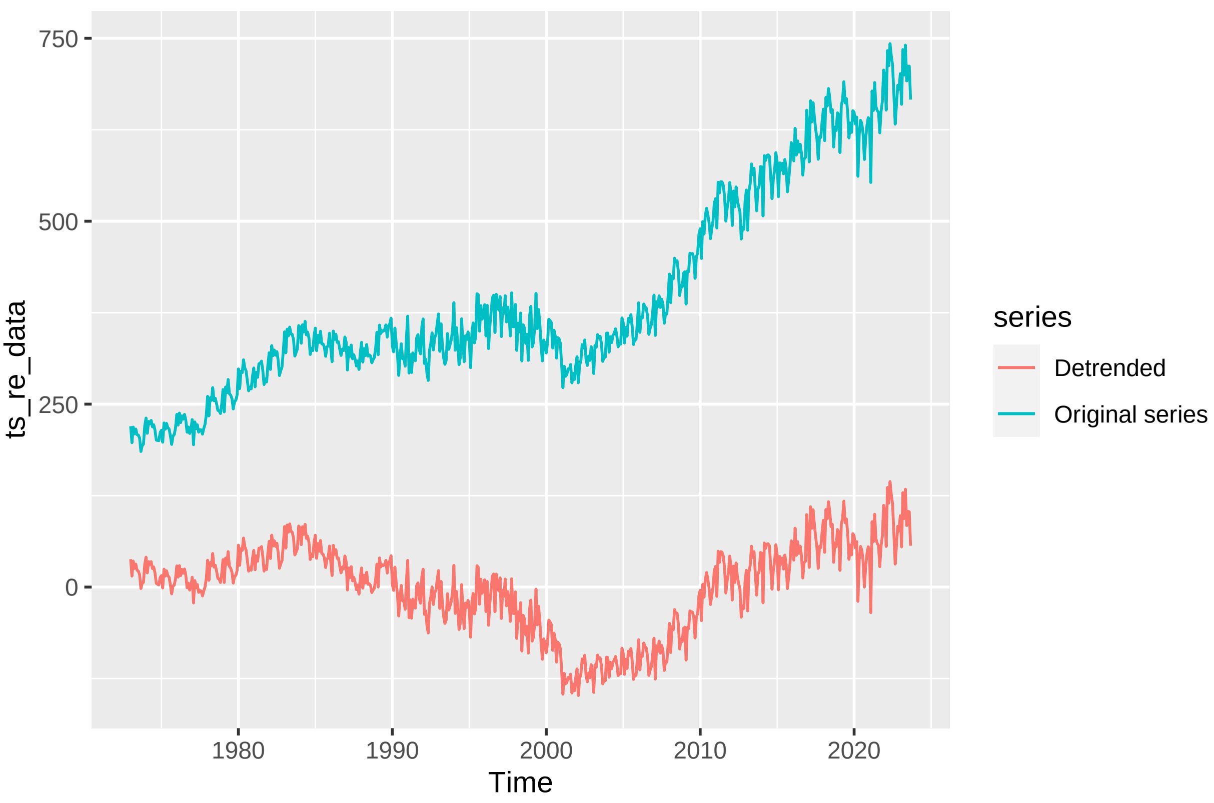
Now let's compare the differenced series with the detrended series you calculated on A3. In other words, for the "Total Renewable Energy Production" compare the differenced series from Q1 with the series you detrended in Q2 using linear regression.

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.

```
ts_dt_re_linear <- ts(dt_re_linear,frequency=12, start = c(1973,1))
head(ts_dt_re_linear)
```

```
##           Jan      Feb      Mar      Apr      May      Jun
## 1973 38.14556 14.93252 35.58448 25.52444 31.47240 23.03536
```

```
autoplot(ts_re_data,series="Original series")+ autolayer(dt_re_linear,series="Detrended") #+autolayer(d
```



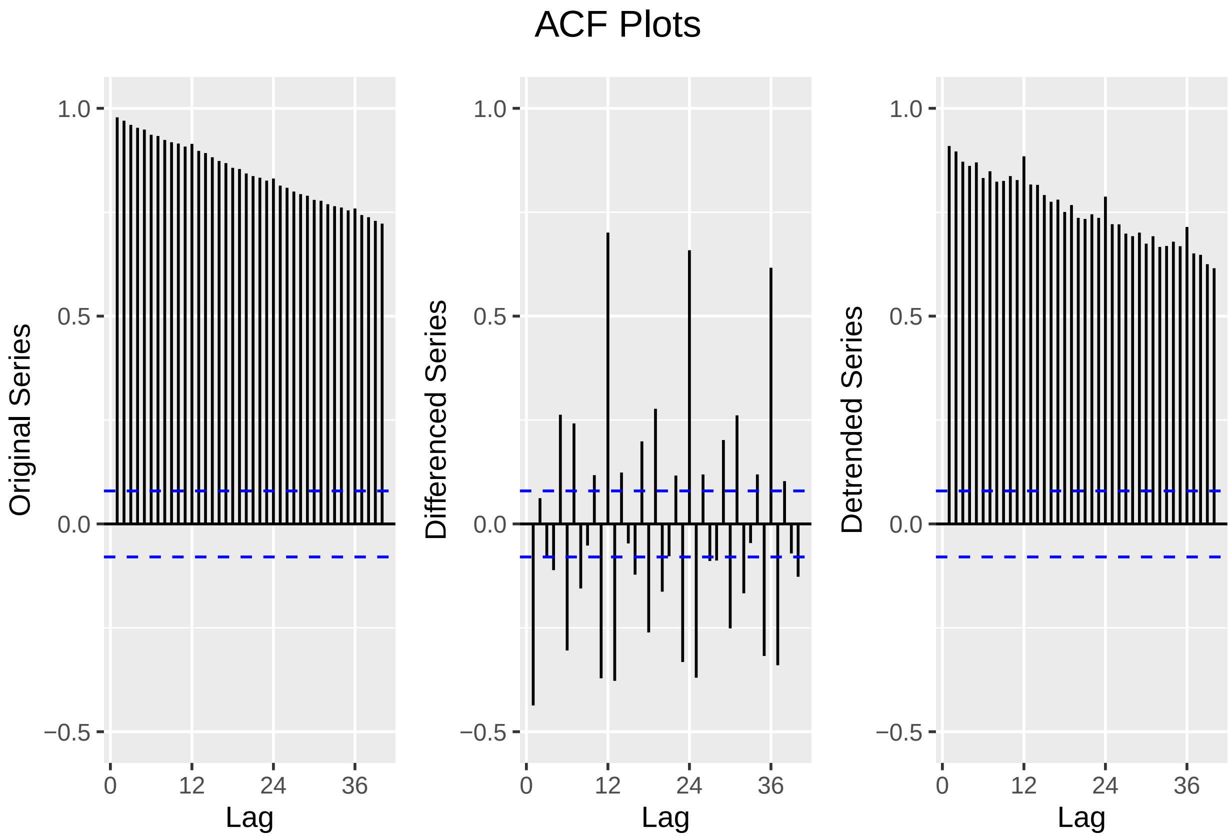
Q4

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?

```
p4_title<-ggdraw()+draw_label("ACF Plots")
p4_plot_row <- plot_grid(
  autoplot(Acf(ts_re_data,lag.max=40,plot=FALSE),main=NULL,ylim=c(-0.5,1),ylab = "Original Series"),
  autoplot(Acf(diff_re_data,lag.max=40,plot=FALSE),main=NULL,ylim=c(-0.5,1),ylab = "Differenced Series"
  autoplot(Acf(ts_dt_re_linear,lag.max=40,plot=FALSE),main=NULL,ylim=c(-0.5,1),ylab = "Detrended Series
  nrow=1,ncol=3
)
```

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

```
p4 <- plot_grid(p4_title,p4_plot_row,nrow=2,ncol=1,rel_heights = c(0.1,1))
p4
```



The differencing was more efficient at eliminating the trend than the linear regression.

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 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.

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

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

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

```
## 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_re_data,alternative = "stationary"))
```

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

Conclusion from SMK test: The S is positive so there is an increasing trend. The p value is <0.05 so the null hypothesis is rejected and we have a significant trend in the series.

Conclusion from ADF test: The p-value is >0.05 so we cannot reject the null hypothesis. The data has a unit root and 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().

```r
re_data_matrix <- matrix(ts_re_data,byrow=FALSE,nrow=12)
```

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

```r
re_data_yearly <- colMeans(re_data_matrix)
head(re_data_yearly)
```

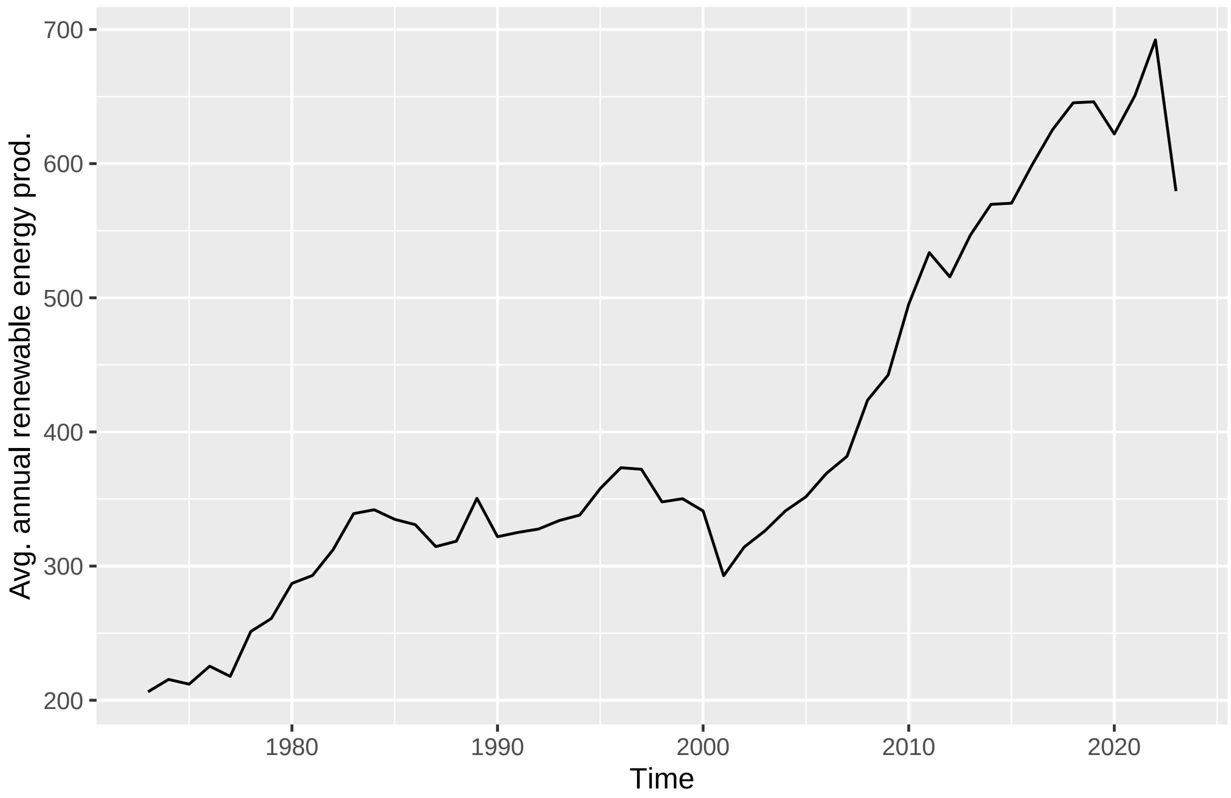
```
## [1] 206.2953 215.5001 212.0139 225.3914 217.7895 251.2457
```

```r
my_year <- c(year(first(energy_date)):year(last(energy_date)))

re_data_new_yearly <- data.frame(my_year, re_data_yearly)

ts_re_yearly <- ts(re_data_yearly, start = c(1973))

p6<-autoplot(ts_re_yearly,ylab = "Avg. annual renewable energy prod.", xlab = "Time")
p6
```



Q7

Apply the Mann Kendal, 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?

```
print("Results of Mann Kendall on average yearly series")
```

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

```
print(summary(MannKendall(ts_re_yearly)))
```

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

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

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

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

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

```r
print("Results from Spearman Correlation")
```

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

```r
sp_rho=cor(ts_re_yearly,my_year,method="spearman")
print(sp_rho)
```

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

```r
#with cor.test you can get test statistics
sp_rho=cor.test(ts_re_yearly,my_year,method="spearman")
print(sp_rho)
```

```
##
##  Spearman's rank correlation rho
##
## data:  ts_re_yearly and my_year
## S = 1908, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9136652
```

The results from the annual average series are in agreement with those from the monthly series.

Mann-Kendall - Series has a trend because p-value is <-0.05. Spearman - the rho is 0.91 i.e. very close to 1. So, series has a linear trend. ADF - Unable to reject null hypothesis because p-value is >0.05 so series has stochastic trend.