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

Assignment 4 - Due date 02/11/25

Rachael Stephan

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.\

```r
#Load/install required package here
library(openxlsx); library(ggplot2); library(forecast);
library(tidyverse); library(tseries); library(Kendall);
library(cowplot); library(lubridate); library(knitr)
```

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 - you may copy your code from A3
#set theme
mytheme <- theme_bw(base_size = 10)+
  theme(axis.title = element_text(size = 10, hjust = 0.5),
        plot.title.position = "panel",
        panel.border = element_rect(colour = "black", fill = NA, linewidth = 0.25),
        plot.caption = element_text(hjust = 0),
        legend.box = "vertical",
        legend.location = "plot",
        axis.gridlines = element_line(color = "grey", linewidth = 0.25),
        axis.ticks = element_line(color = "black", linewidth = 0.5),
        axis.grid = element_blank())
theme_set(mytheme)

#upload dataset
renewable_e_prod_consump <-
  read.xlsx("./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
                                      sheet = "Monthly Data",
                                      startRow = 13,
                                      colNames = FALSE)

#get column names
col_units <-
  read.xlsx("./Data/Table_10.1_Renewable_Energy_Production_and_Consumption_by_Source.xlsx",
            rows = 11:12,
            sheet="Monthly Data",
            colNames=FALSE)

#set col names
colnames(renewable_e_prod_consump) <- col_units[1,]

#fix dates
renewable_e_prod_consump$Month <- as_date(renewable_e_prod_consump$Month, origin = "1900-01-01")
renewable_e_prod_consump$Month <- paste(month(renewable_e_prod_consump$Month,
                                              label = TRUE,
                                              abbr = TRUE),
                                        year(renewable_e_prod_consump$Month))

#select for columns of interest
renewable_matrix <- renewable_e_prod_consump %>%
  select(`Month`,
         `Total Renewable Energy Production`)

#get first few rows of each column to check structure and values
kable(head(renewable_matrix),
      caption = "First few rows of the selected timeseries for analysis")
```

Table 1: First few rows of the selected timeseries for analysis

Month	Total Renewable Energy Production
Jan 1973	219.839
Feb 1973	197.330
Mar 1973	218.686
Apr 1973	209.330
May 1973	215.982
Jun 1973	208.249

```
str(renewable_matrix)
```

```
## ’data.frame’:    621 obs. of  2 variables:
```

```
##  $ Month                            : chr  "Jan 1973" "Feb 1973" "Mar 1973" "Apr 1973" ...
##  $ Total Renewable Energy Production: num  220 197 219 209 216 ...
```

```
#create time series object
ts_renewable <- ts(renewable_matrix[,2],
                start=c(1973,1),
                frequency=12)
```

Stochastic Trend and Stationarity Tests

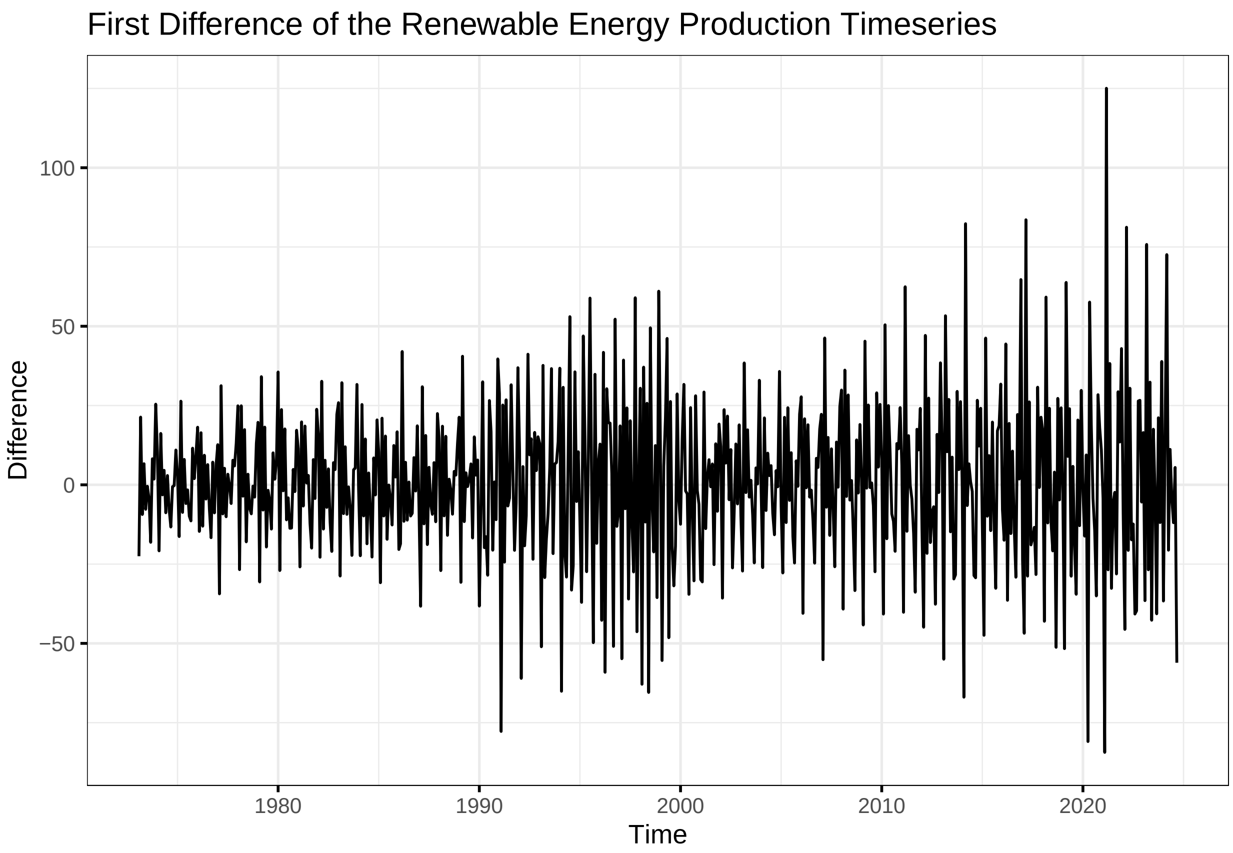
For this part you will work only with the column Total Renewable Energy Production.

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?

```
#difference the renewable energy timeseries
ts_renewable_diff <- diff(ts_renewable,
                       lag = 1,
                       differences = 1)

#plot the differenced timeseries
autoplot(ts_renewable_diff)+
  labs(title = "First Difference of the Renewable Energy Production Timeseries",
       y = "Difference")
```


First Difference of the Renewable Energy Production Timeseries

The series does not appear to have a linear trend over time (i.e., there does not appear to be a slope). However, the magnitude of the differences appear to increase in both directions over time, creating a slight funnel shape.

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.

```r
time <- 1:length(ts_renewable)

#Fit the linear trend for renewable energy
lm_renewable <- lm(ts_renewable ~ time)

#save renewable lm coefficients
lm_renewable_beta0 <- as.numeric(lm_renewable$coefficients[1])
lm_renewable_beta1 <- as.numeric(lm_renewable$coefficients[2])

#create renewable linear trend equation
lm_trend_renewable <- lm_renewable_beta0 + lm_renewable_beta1 * time

#use linear trend to detrend renewable ts
lm_detrend_renewable <- ts_renewable - lm_trend_renewable
```

```r
#create detrended timeseries
ts_renewable_detrend <- ts(lm_detrend_renewable,
                              start = c(1973,1),
                              frequency = 12)
```

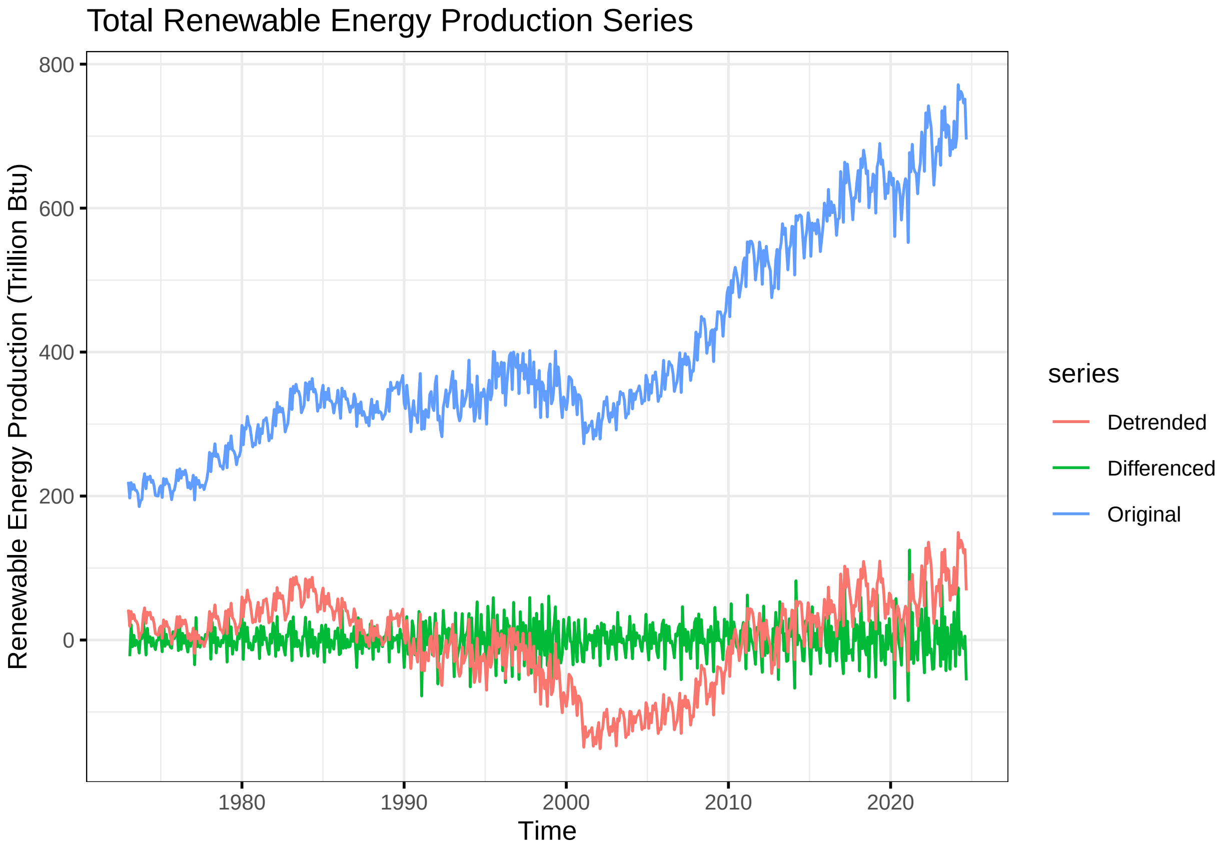
Q3

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?

```r
#create plot
autoplot(ts_renewable, series = "Original")+
  autolayer(ts_renewable_diff, series = "Differenced")+
  autolayer(ts_renewable_detrend, series = "Detrended")+
  labs(title = "Total Renewable Energy Production Series",
       y = paste("Renewable Energy Production",col_units[2,6], sep = " "))
```

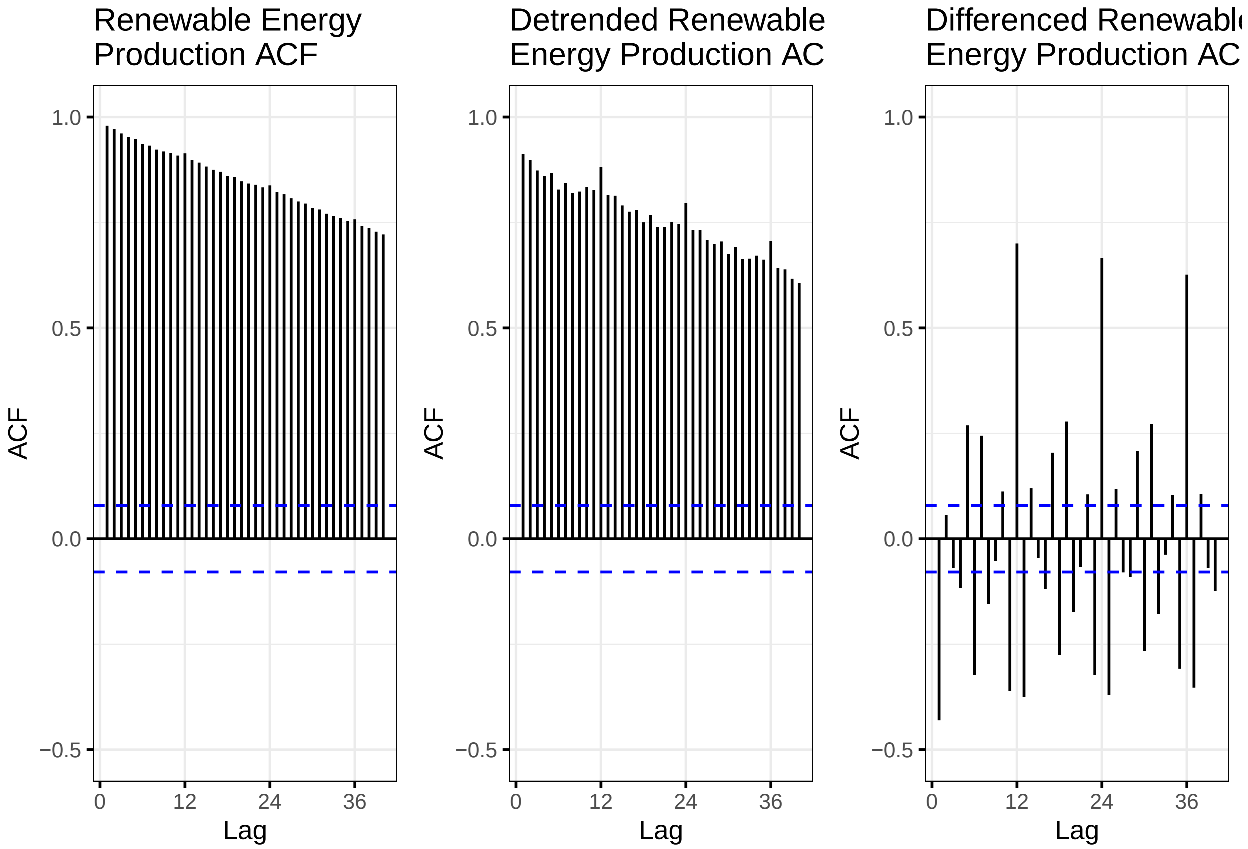


The differenced method appears more efficient in removing the trend. The detrended series still shows a non-linear trend as it fluctuates above and below 0. The differenced series does not appear to trend away from 0, although the magnitudes of the differences (both positive and negative) appear to change over time. Since differencing is used to remove stochastic trends, it is possible that the original series is stochastic over stationary.

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?

```
#calculate the ACF of each series
acf_renewable <- Acf(ts_renewable,
                     lag.max=40,
                     type="correlation",
                     plot=FALSE)

acf_renewable_detrend <- Acf(ts_renewable_detrend,
                     lag.max=40,
                     type="correlation",
                     plot=FALSE)

acf_renewable_diff <- Acf(ts_renewable_diff,
                     lag.max=40,
                     type="correlation",
                     plot=FALSE)

#create an ACF plot for each of the series
acf_plot_renewable <- autoplot(acf_renewable)+
  labs(title = "Renewable Energy\nProduction ACF")+
  ylim(c(-0.5,1))

acf_plot_renewable_detrend <- autoplot(acf_renewable_detrend)+
  labs(title = "Detrended Renewable\nEnergy Production ACF")+
  ylim(c(-0.5,1))

acf_plot_renewable_diff <- autoplot(acf_renewable_diff)+
  labs(title = "Differenced Renewable\nEnergy Production ACF")+
  ylim(c(-0.5,1))

#plot ACF comparisons in a plot
plot_grid(acf_plot_renewable,
          acf_plot_renewable_detrend,
          acf_plot_renewable_diff,
          align = "h",
          nrow = 1)
```



Answer: The differenced renewable energy ACF plot shows more evidence of this method being effective at eliminating the trend. The slow decay across increasing lags in the original plot indicates the presence of a strong trend. In the detrended data, the decay may be slightly faster but it is still slow, and all of the lags shown are significant. There is some presence of seasonality shown by oscillations in the decaying trend. The differenced ACF plot has a quicker decay, fewer significant lags, and smaller correlation magnitudes than the detrended plot. This indicates that it was more effective at removing the trend. There are still significant lags in the differenced plot. However, given the repetitive pattern of which lags are significant and the sign of the ACF, these significant lags may also be related to seasonality.

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.

```
#run the seasonal man kendall test
smk_renewable <- SeasonalMannKendall(ts_renewable)

#read the smk output
summary(smk_renewable)
```

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

> The SMK test tests for trends that occur overtime, taking into account the presence of seasonality. The test resulted in a relatively large s value (12468) relative to the denominator (15758.5). This means that successive values in this timeseries had the tendency to get larger. The tau indicates the magnitude and sign of the relationship. A tau of 0.791 indicates a strong positive trend over time. The p-value for this test was =< 2.22e^{-16}. Therefore, the results of this SMK test are significant.

```
#run the augmented Dickey-Fuller test
adf_renewable <- adf.test(ts_renewable,
                          alternative = "stationary")

#read the adf output
adf_renewable
```

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

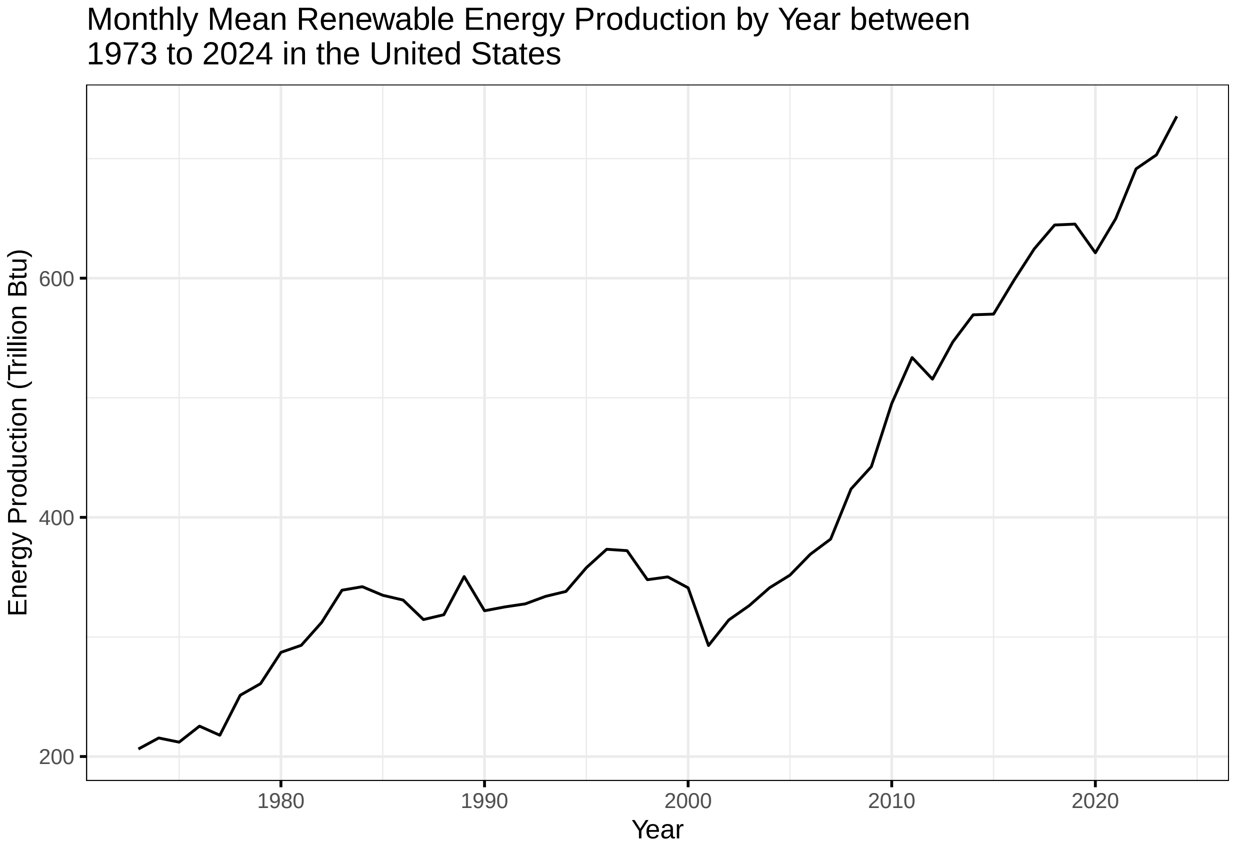
> The augmented Dickey-Fuller test is examining whether a timeseries is stationary (i.e., properties, like the mean, do not change over time) or stochasitic (i.e., properties change over time). The alternative hypothesis of the ADF is that the renewable energy production timeseries is stationary. The p-value of this test is 0.9242. This means we fail to reject the null hypothesis, and we conclude that the renewable energy production time series has a unit root and is stochastic over a stationary trend. This matches what the original trend shows in Q3. The original trend goes through a perid of general increase, then it stabilizes before and general increase again. Differencing is also used to remove a stochastic trend, and the differenced series was more effective at removing the trend compared to the detrended series. This indicates that the properties of the time series are not constant (i.e., 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 accumulates yearly series into a time series object and plot the series using autoplot().

```
#Group data by month (row) and year (column)
renewable_matrix_yearly <- matrix(ts_renewable,
                                  byrow=FALSE,
                                  nrow=12,
                                  dimnames = list(c("Jan", "Feb", "Mar", "Apr",
                                                    "May","June", "July", "Aug",
                                                    "Sept","Oct", "Nov", "Dec"),
                                                  c(1973:2024)))
```

```
#the last entry is sept 2024. So reset obs for oct - dec. to NA
renewable_matrix_yearly[10:12, 52] <- NA

#set the years this ts occurs over
my_year <- c(1973:2024)

#get the means of each year and put into a dataframe
renewable_matrix_yearly_mean <- colMeans(renewable_matrix_yearly,
                                    na.rm = TRUE)
renewable_matrix_yearly_mean <- data.frame("Year" = my_year,
                                           "renewable_energy" = renewable_matrix_yearly_mean)

#create a timeseries for the yearly mean dataset
ts_renewable_yearly <- ts(renewable_matrix_yearly_mean[,2],
                          start = c(1973),
                          frequency = 1)

#plot the yearly timeseries
autoplot(ts_renewable_yearly)+
  labs(title = "Monthly Mean Renewable Energy Production by Year between\n1973 to 2024 in the United Sta
       y = paste("Energy Production",col_units[2,6], sep = " "),
       x = "Year")
```



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?

```r
#run the seasonal man kendall test
mk_renewable_yearly <- MannKendall(ts_renewable_yearly)

#read the smk output
summary(mk_renewable_yearly)
```

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

> Answer: The s value between the two series are not comparable due to a different denominator. The tau for the monthly mean renewable energy production by year timeseries ("yearly series") was 0.817. This is larger than the tau that was produced for the original renewable energy timeseries (0.791). This indicates that the yearly series also has a strong positive trend over time that is more prominent than the original series. However, the values are relatively similar, which indicates that the trend observed is similar. The p-value for this yearly timeseries Mann-Kendall test was =< $2.22e^{-16}$. Therefore, the results for this timeseries are also significant, and these tests are in agreement.

```r
#run the Spearman correlation rank test
spearman_renewable_yearly <- cor.test(ts_renewable_yearly,my_year,method="spearman")

#read the Spearman output
spearman_renewable_yearly
```

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

> Answer: The Spearman correlation rank test was not conducted on the original series because it cannot handle seasonality. Since the yearly series has averaged the values across a year, seasonality will not be present in the yearly series. The Spearman rank correlation test is looking for monotonic relationships between two variables. The rho value is 0.9209425, which indicates a strong, positive monotonic trend. The p-value is $< 2.2e^{-16}$. Therefore, the results for this timeseries are also significant. This test is in agreement with both of the Mann-Kendall tests, for the original and yearly timeseries.

```r
#get adf for yearly data
adf_renewable_yearly <- adf.test(ts_renewable_yearly,
                                 alternative = "stationary")

#read the adf output
adf_renewable_yearly
```

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

> Answer: Both the yearly series and original series had p-values larger than alpha (p-value > 0.05). The alternative hypothesis of the ADF is that the yearly renewable energy production timeseries is stationary. Therefore, we fail to reject the null hypothesis, and we conclude that the yearly time series has a unit root and is stochastic (i.e., nonstationary) over a stationary trend. This matches the ADF results for the original monthly series.