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

Assignment 4 - Due date 02/11/25

Chloe Young

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.\

```
library(openxlsx)
library(readxl)
library(lubridate)
library(ggplot2)
library(forecast)
library(Kendall)
library(tseries)
library(trend)
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"**.

```
#Importing data set - you may copy your code from A3

#Importing data set without change the original file using read.xlsx
energy_data1 <- read_excel(path="/home/guest/TSA_Sp25/Data/Table_10.1_Renewable_Energy_Production_and_C
```

```r
#Now let's extract the column names from row 11
read_col_names <- read_excel(path="/home/guest/TSA_Sp25/Data/Table_10.1_Renewable_Energy_Production_and_

#Assign the column names to the data set
colnames(energy_data1) <- read_col_names

energy_data1 <- energy_data1[ ,5] #the space before the comma means you want just column 5

#Visualize the first rows of the data set
head(energy_data1)
```

```
## # A tibble: 6 x 1
##   `Total Renewable Energy Production`
##                                 <dbl>
## 1                                220.
## 2                                197.
## 3                                219.
## 4                                209.
## 5                                216.
## 6                                208.
```

```r
nobs <- nrow(energy_data1)
```

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.

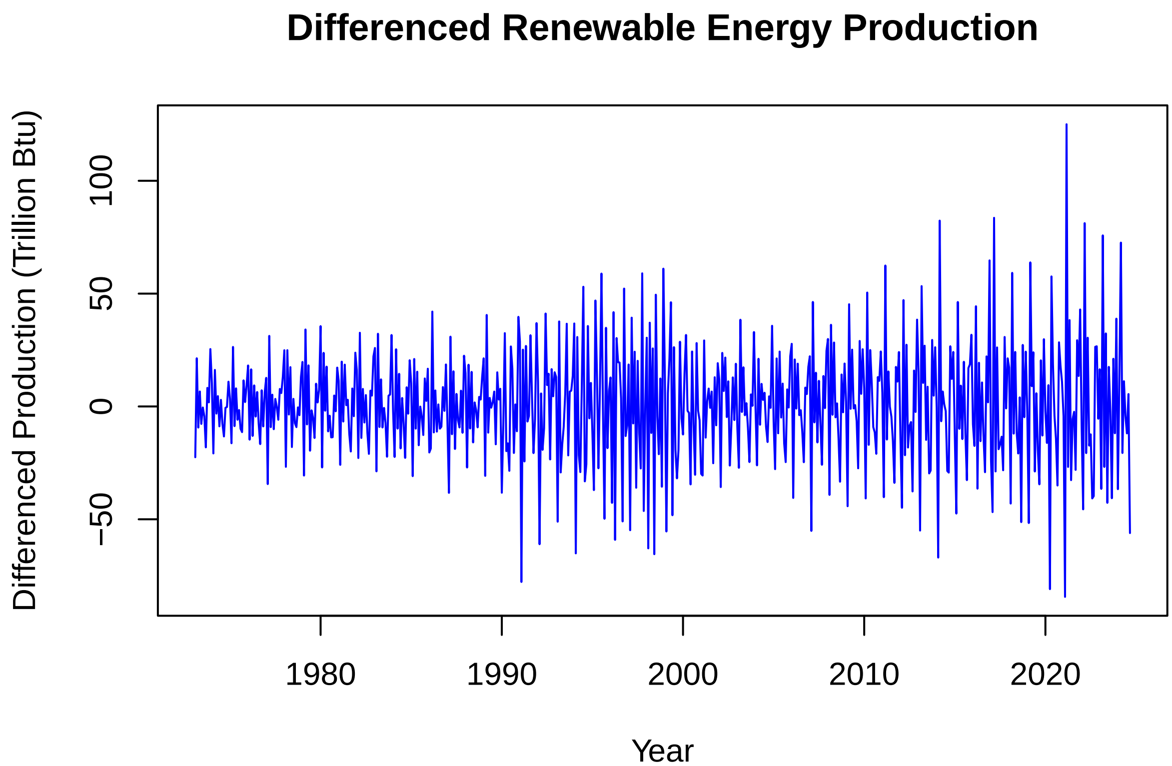
Try differencing at lag 1 only once, i.e., make `lag=1` and `differences=1`. Plot the differenced series. Do the series still seem to have trend?

```r
energy_data1 <- as.numeric(unlist(energy_data1))

diff_renewable <- diff(energy_data1, lag = 1, differences = 1)

diff_ts_renewable <- ts(diff_renewable, start=c(1973,2), frequency=12)

plot(diff_ts_renewable, type="l", col="blue", main="Differenced Renewable Energy Production",
     xlab="Year", ylab="Differenced Production (Trillion Btu)")
```



The plot shows an overall increasing trend as the differentiated production gets larger with time, so the more months pass, the more differentiated renewable production becomes.Therefore, the series does appear to still have a trend, although it looks weaker than it does in the original graph.

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.

```
#regression
ts_renewable <- ts(energy_data1, start=c(1973,1),frequency=12)

#create vector t
t <- c(1:nobs)

#combine t and original into one data frame
data_r <- data.frame("t"=t,"ts_data_r"=ts_renewable)

#Fit a linear trend to TS, lm function needs a data frame object
linear_model_r=lm(ts_data_r~t,data_r)

summary(linear_model_r)
```

```
##
```

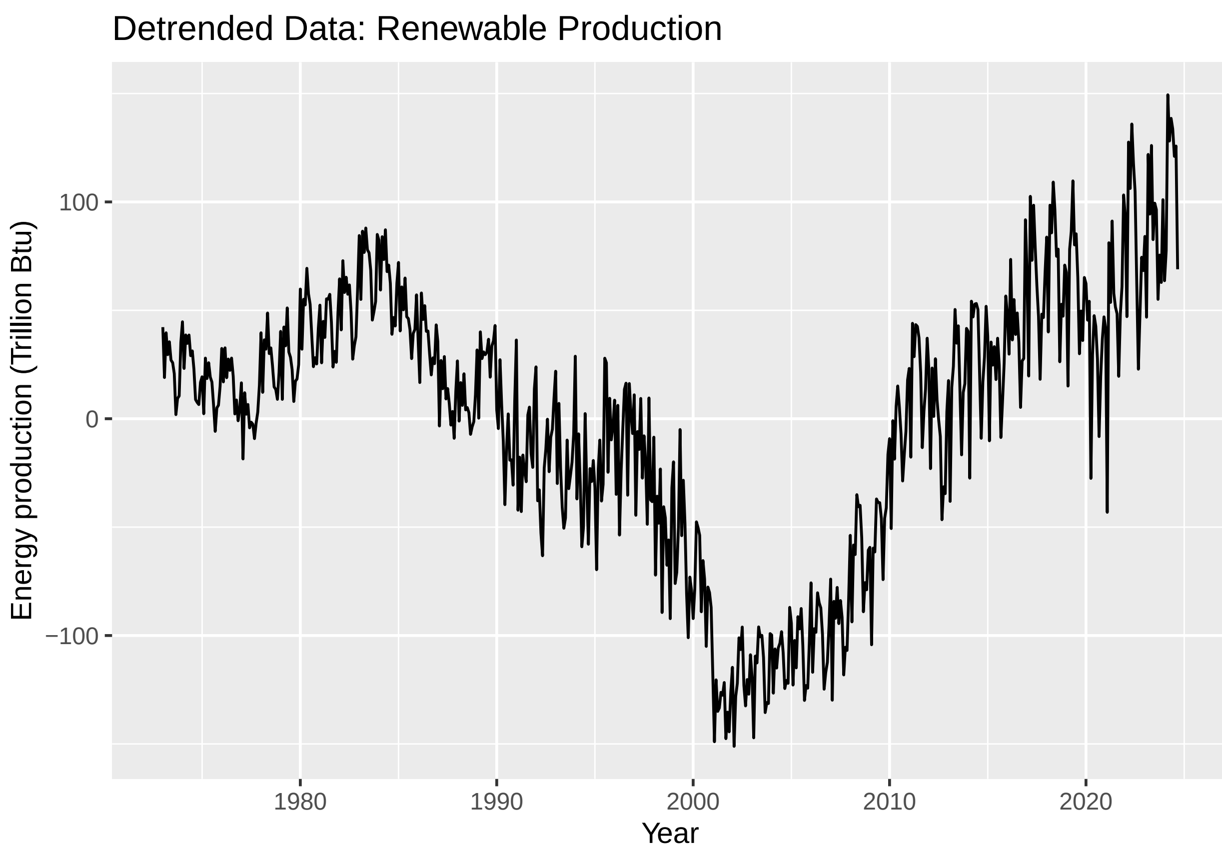
```
## Call:
## lm(formula = ts_data_r ~ t, data = data_r)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -151.11  -37.84   13.53   41.76  149.42
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 176.87293    4.96189   35.65   <2e-16 ***
## t             0.72393    0.01382   52.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61.75 on 619 degrees of freedom
## Multiple R-squared:  0.8159, Adjusted R-squared:  0.8156
## F-statistic:  2743 on 1 and 619 DF,  p-value: < 2.2e-16
```

```r
#Detrend Renewable
beta0_r <- as.numeric(linear_model_r$coefficients[1])
beta1_r <- as.numeric(linear_model_r$coefficients[2])

linear_trend_r <- beta0_r + beta1_r * t
ts_linear_r <- ts(linear_trend_r,start=c(1973,1),frequency=12)

detrend_data_r <- ts_renewable - linear_trend_r
ts_r_detrend <- ts(detrend_data_r, start=c(1973,1),frequency = 12)

#Plot
detrend_r_plot <- autoplot(ts_r_detrend) +
  ggtitle("Detrended Data: Renewable Production")+
  xlab("Year")+
  ylab("Energy production (Trillion Btu)")
detrend_r_plot
```



Q3

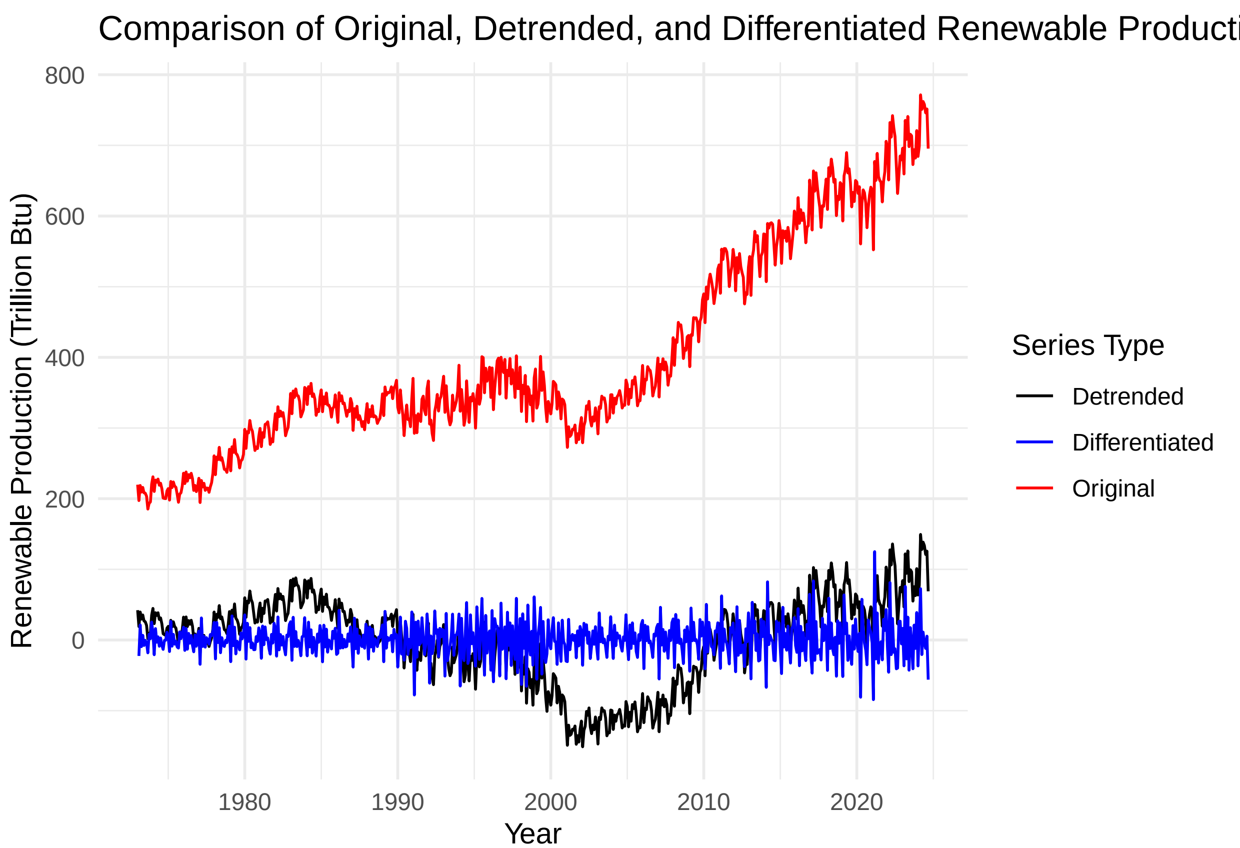
Now let's compare the differenced series with the detrended series you calculated on A3. In other words, for the "Total Renewable Energy Production" compare the differenced series from Q1 with the series you detrended in Q2 using linear regression.

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?

```
compare_plot <- autoplot(ts_renewable, series="Original") +
  autolayer(ts_r_detrend, series="Detrended") +
  autolayer(diff_ts_renewable, series="Differentiated") +
  xlab("Year") +
  ylab("Renewable Production (Trillion Btu)") +
  labs(color="Series Type") +
  ggtitle("Comparison of Original, Detrended, and Differentiated Renewable Production") +
   scale_color_manual(values = c("black", "blue", "red")) +
  theme_minimal()

compare_plot
```



Answer:

The original series shows an increasing trend. Although the detrended series somewhat fluctuates around zero, there is somewhat of an increasing pattern still visible. Similarly, with the differentiated series, there is a subtle byt visible increasing. The differentiated series was the best at removing the trend of the methods, but neither did it completely successfully.

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?

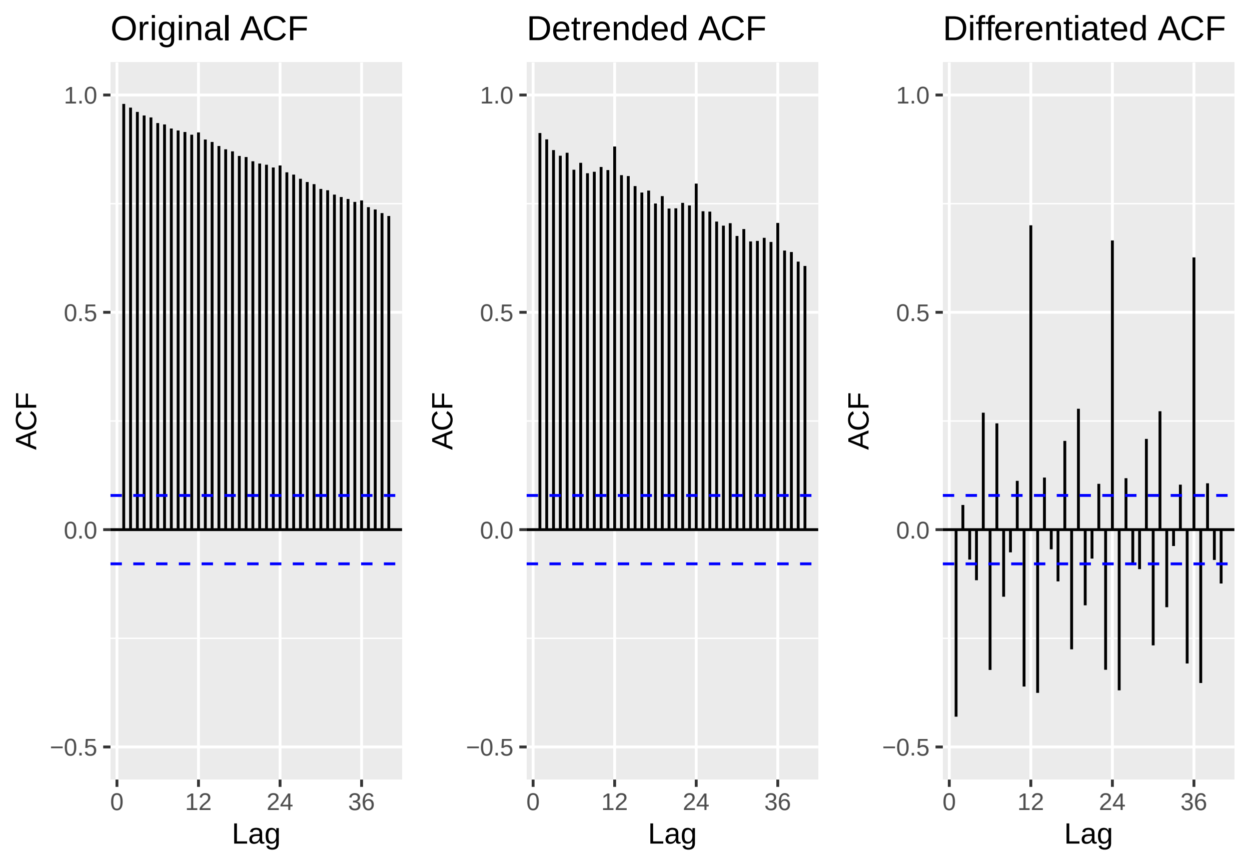
```
#plot ACF
OriginalRenewable_acf=ggAcf(ts_renewable, lag.max=40, type="correlation", plot=TRUE) +
  ggtitle("Original ACF") +
  coord_cartesian(ylim=c(-0.5,1))

DetrendRenewable_acf=ggAcf(ts_r_detrend, lag.max=40, type="correlation", plot=TRUE) +
  ggtitle("Detrended ACF") +
  coord_cartesian(ylim=c(-0.5,1))

DifferentiatedRenewable_acf=ggAcf(diff_ts_renewable, lag.max=40, type="correlation",  plot=TRUE) +
```

```r
  ggtitle("Differentiated ACF") +
  coord_cartesian(ylim=c(-0.5,1))

plot_grid(OriginalRenewable_acf, DetrendRenewable_acf, DifferentiatedRenewable_acf, ncol = 3)
```



Answer:

The most effective method for removing the trend was differentiating the series. Both the original and detrended plots show a similar pattern of a gradual decrease, displaying long term dependence in the data, or a trend. The differentiated option, however, shows more variety in the plot with some statistically unsignificant spikes, indicating that it did a better job of removing the trend.

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
print("Result from Mann Kendall test")
```

```
## [1] "Result from Mann Kendall test"
```

```r
seasonal_mk_test <- MannKendall(ts_renewable)
print(seasonal_mk_test)
```

```
## tau = 0.759, 2-sided pvalue =< 2.22e-16
```

```r
#ADF Test
print("Result from ADF test")
```

```
## [1] "Result from ADF test"
```

```r
adf_test <- adf.test(ts_renewable, alternative = "stationary")
print(adf_test)
```

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

Answer:

Mann Kendall Interpretation: The p-value from the Mann Kendall test is 2.22e-16, which is less than 0.05. Therefore, the null hypothesis is rejected and it can be said that there is a statistically significant monotonic trend in the data. Since tau is 0.759, this indicates that the trend is increasing (since the value is positive) and relatively strong (since the value is relatively close to 1).

ADF Interpretation: The p-value from the ADF test is 0.9242 which is considerably greater than 0.05, meaning that the null hypothesis isn't rejected, so the series is non-stationary and has a stochastic trend.

The results don't match what previous questions showed since the differentiated graph still had a trend, although it was the best at removing the trend compared to the detrending option, so in that sense the results somewhat matched since having a stochastic trend would indicate that differentiating was successful at 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
#Group data in yearly steps instances
renewable_data_matrix <- matrix(ts_renewable,byrow=FALSE,nrow=12)
```

```
## Warning in matrix(ts_renewable, byrow = FALSE, nrow = 12): data length [621] is
## not a sub-multiple or multiple of the number of rows [12]
```

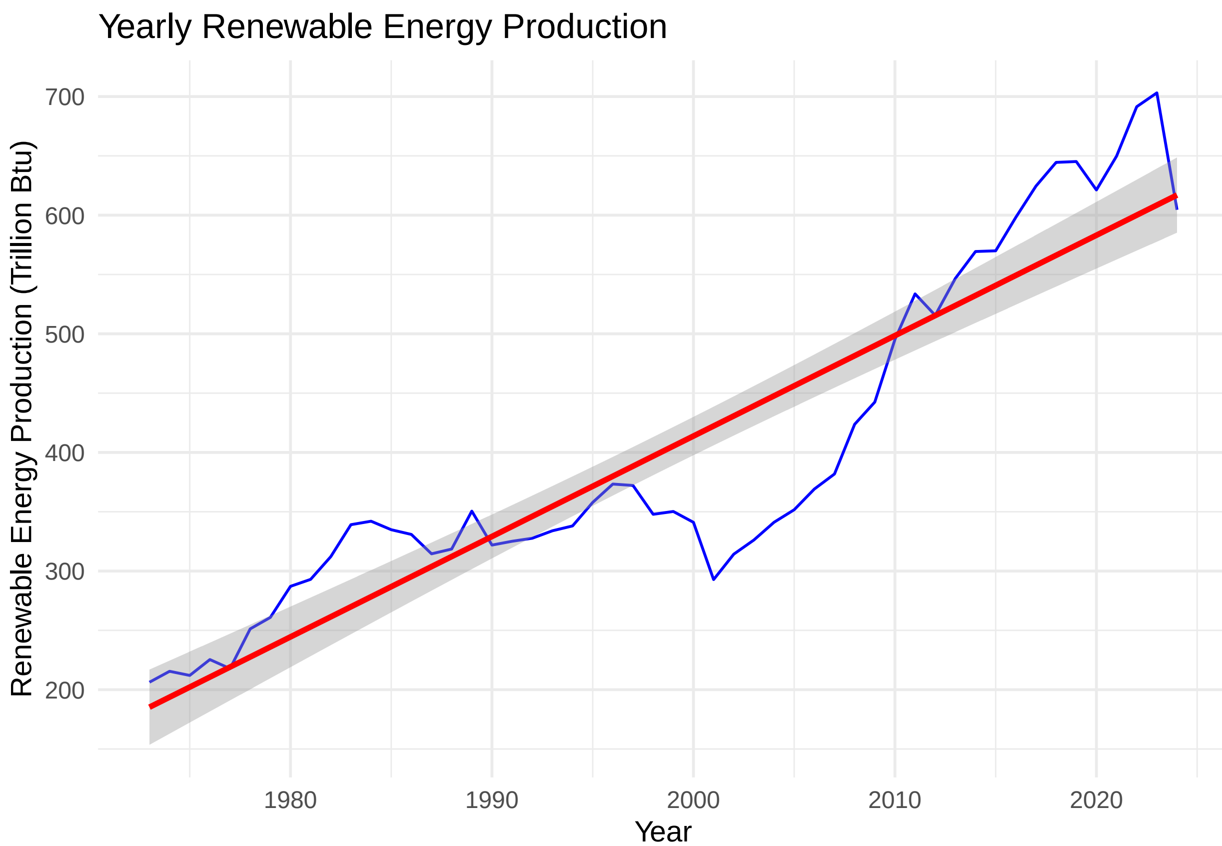
```r
renewable_data_yearly <- colMeans(renewable_data_matrix)

my_year <- c(1973:2024)

renewable_data_yearly <- data.frame(my_year, "renewable_data"=renewable_data_yearly)

ggplot(renewable_data_yearly, aes(x=my_year,y=renewable_data))+
  geom_line(color="blue")+
  geom_smooth(color="red",method="lm") +
  labs(
    title="Yearly Renewable Energy Production",
    x="Year",
    y="Renewable Energy Production (Trillion Btu)"
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



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
ts_renewable_yearly <- ts(renewable_data_yearly$renewable_data, start=1973, frequency = 1)

#Seasonal Mann-Kendall Test
print("Mann-Kendall test")
```

```
## [1] "Mann-Kendall test"
```

```r
seasonal_mk_test_yearly <- MannKendall(ts_renewable_yearly)
print(seasonal_mk_test_yearly)
```

```
## tau = 0.807, 2-sided pvalue =< 2.22e-16
```

```r
#Spearman Correlation Rank Test
print("Spearman Rank Correlation Test")
```

```
## [1] "Spearman Rank Correlation Test"
```

```r
spearman_test <- cor.test(my_year, renewable_data_yearly$renewable_data, method="spearman")
print(spearman_test)
```

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

```r
#ADF Test
print("ADF test")
```

```
## [1] "ADF test"
```

```r
adf_test_yearly <- adf.test(ts_renewable_yearly, alternative="stationary")
print(adf_test_yearly)
```

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

Answer:

Mann Kendall Interpretation: The p-value from the Mann Kendall test is 2.22e-16, which is less than 0.05. Therefore, the null hypothesis is rejected and it can be said that there is a statistically significant monotonic trend in the data. Since tau is 0.807, this indicates that the trend is increasing (since the value is positive) and relatively strong (since the value is relatively close to 1). This tau value is greater than the one found when testing the monthly data, indicating a stronger yearly trend than monthly trend, likely because there is less variation over the years than months. Nonetheless, the results are in agreement with those found when testing monthly data, as both show a significant, increasing trend.

Spearmans Rank Interpretation: The p-value from the Spearmans Rank test is 2.22e-16, which is less than 0.05, so the data is statistically significant and the null is rejected, meaning that there is a monotonic relationship in the data. The Spearmans Rho value is 0.91 which is very close to one, indicating a strong, positive trend. Therefore, as the years progress, renewable energy production increases.

ADF Interpretation: The p-value from the ADF test is 0.7098 which is considerably greater than 0.05, meaning that we fail to reject the null hypothesis, so the series is non-stationary and has a stochastic trend. This is in agreement with the results found in Q5 for the monthly data.