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

Assignment 4 - Due date 02/12/24

Faustin Kambale

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.\

```
library(lubridate)
library(ggplot2)
library(forecast)
library(Kendall)
library(tseries)
library(cowplot)
library(dplyr)
library(tidyr)
```

Questions

Consider the same data you used for A3 from the spreadsheet "Table_10.1_Renewable_Energy_Production_and_Consumption
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".

```
raw_data <- read.csv(file
="/Users/faustinkambale/Library/CloudStorage/OneDrive-DukeUniversity/Spring 2024 classes/Time Series 4 
                     header=TRUE,skip=0)
data <- raw_data[,c(1,5)] #Working with renewable energy Data set
colnames(data) <- c("date","renen")
data$date <- ym(data$date)
nobs <- nrow(data)
head(data,10)
```

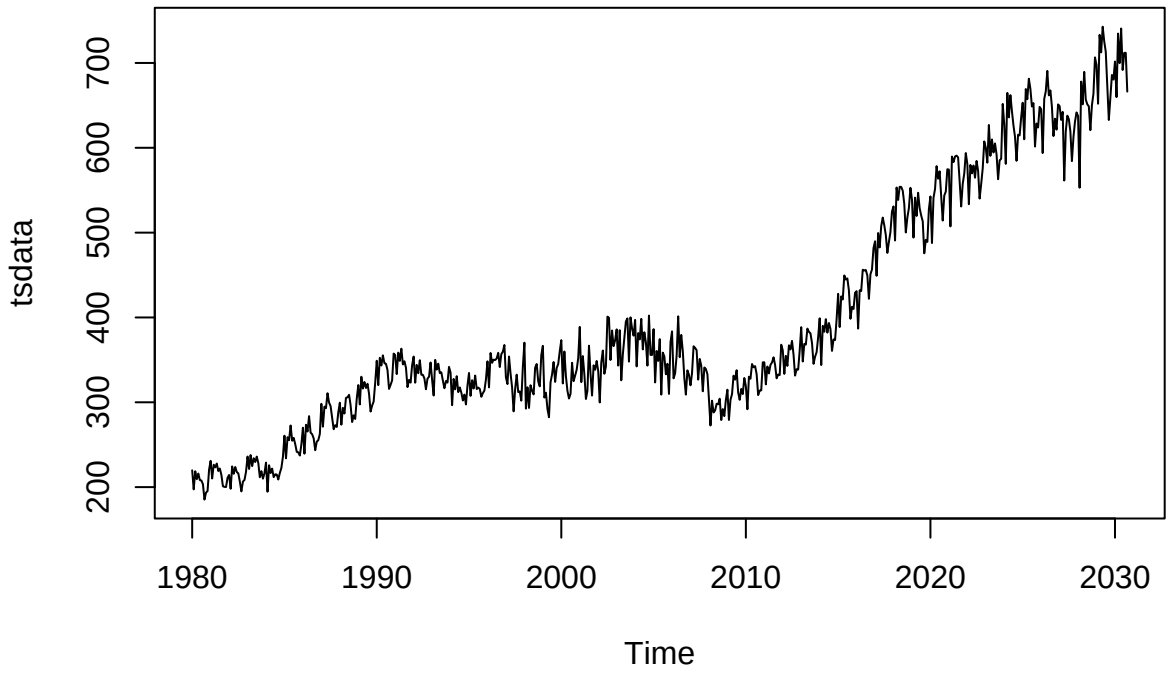
Stochastic Trend and Stationarity Tests

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?

```
#convert in timeseries
tsdata <- ts(data[,2], start=c(1980,1),frequency=12)
plot(tsdata)
```




```
#Differentiating

diff_reneg <- diff(tsdata, lag = 1, differences = 1)  # Adjusting  lag and difference
```

Q2

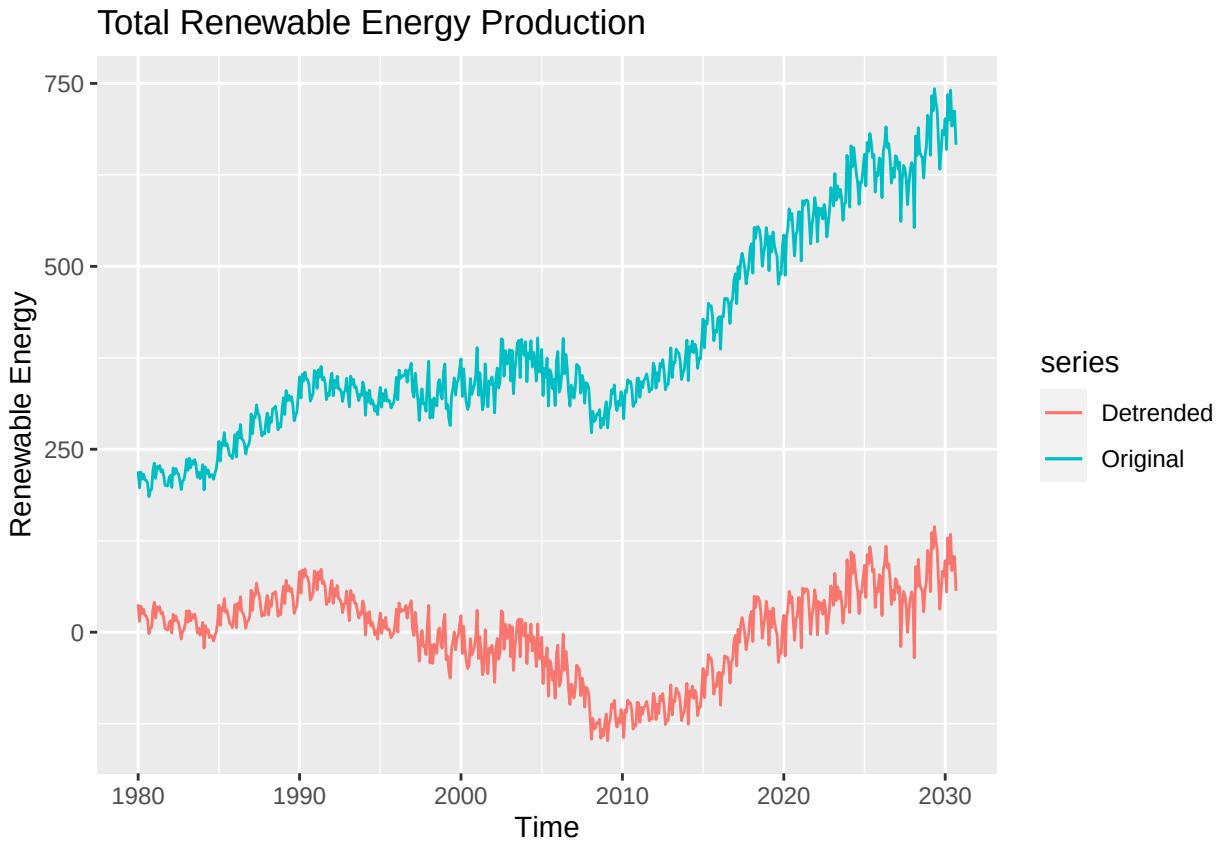
Copy and paste part of your code for A3 where you run the regression for Total Renewable Energy Production and subtract that from the original series. This should be the code for Q3 and Q4. make sure you use the same name for your time series object that you had in A3.

```
t <- 1:nobs #I create vector t
regmodel_renewable=lm(tsdata~t,cbind(tsdata,t)) #regression for renewable energy
beta0_renewable=regmodel_renewable$coefficients[1]
beta1_renewable=regmodel_renewable$coefficients[2]
print(summary(regmodel_renewable))
```

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

```r
#1. Detrend for renewable energy
renewable_detrend <- tsdata - (beta0_renewable+beta1_renewable*t)
renewable_detrend=ts(renewable_detrend, frequency=12,start=c(1980,1))
autoplot(tsdata,series="Original") +
  autolayer(renewable_detrend,series="Detrended") +
  ylab("Renewable Energy") +
  ggtitle("Total Renewable Energy Production")
```

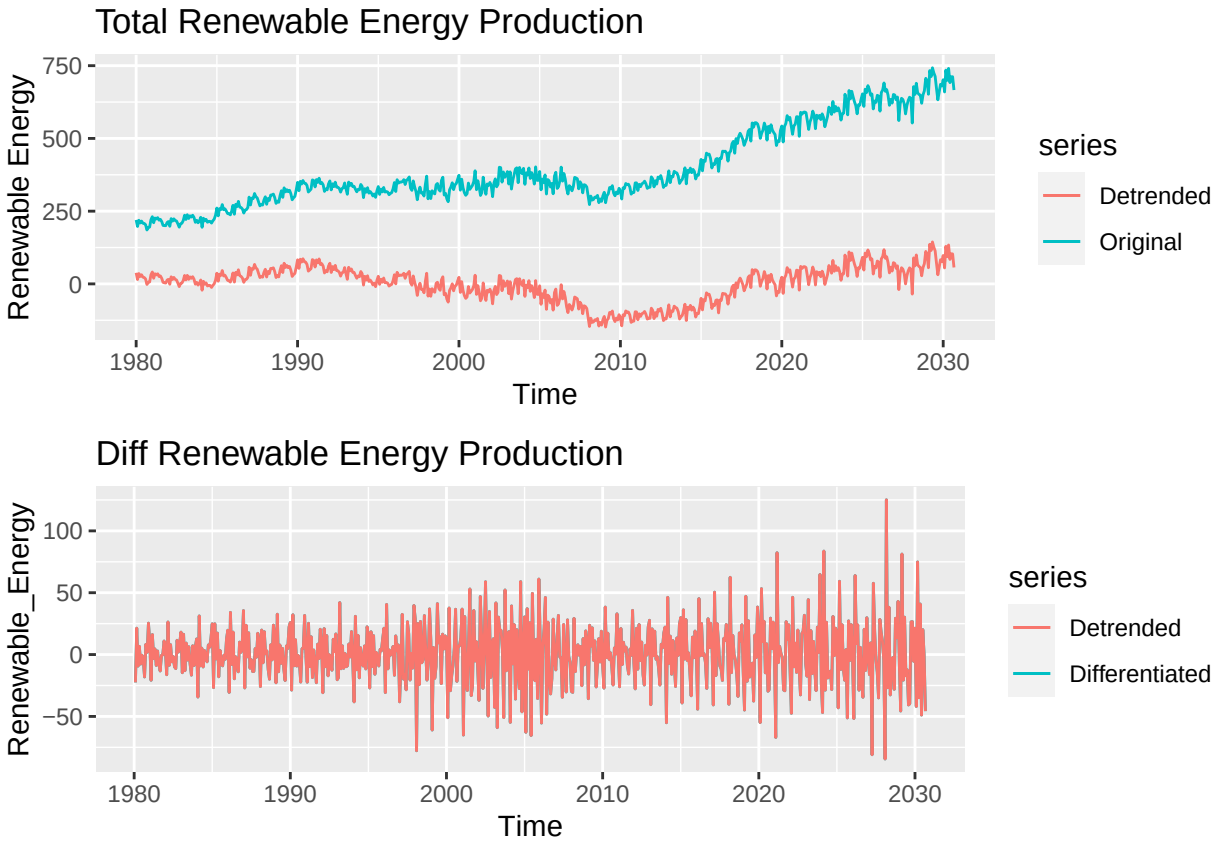


Q3

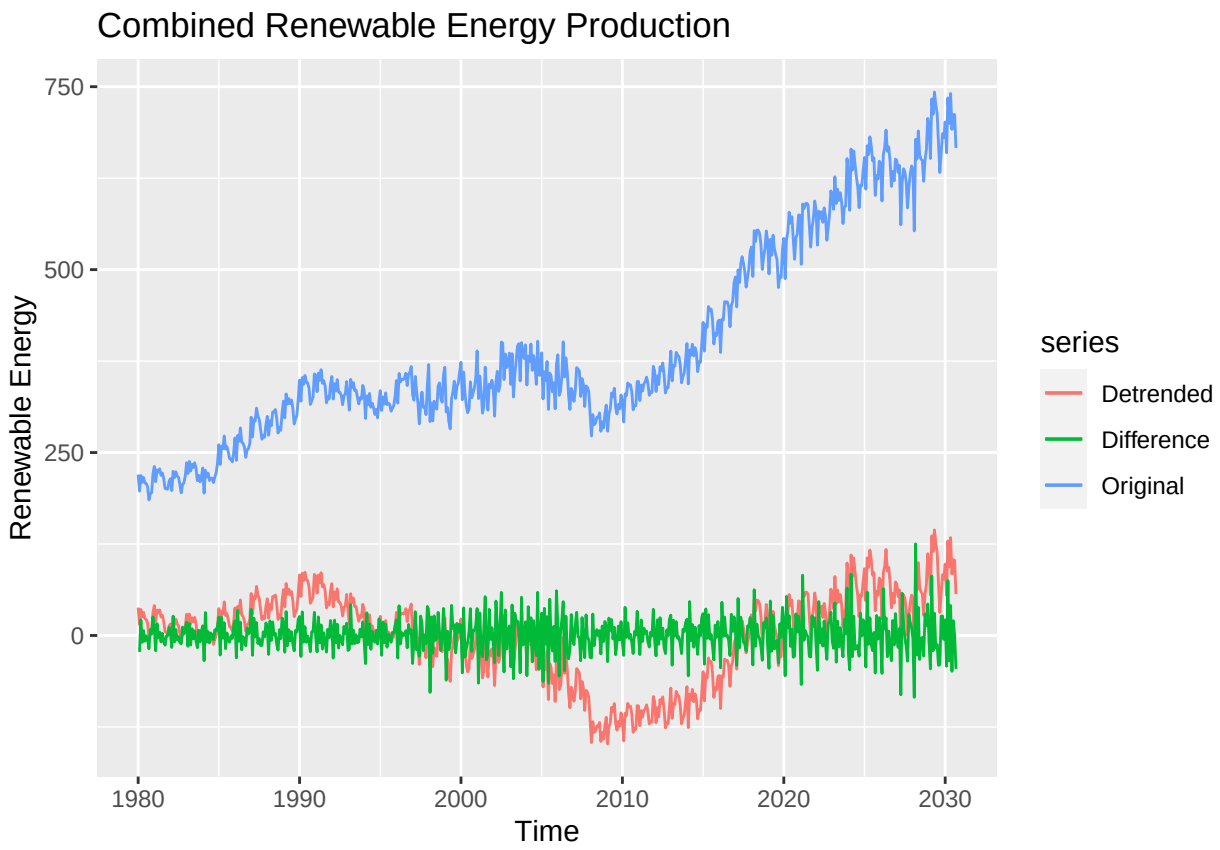
Now let's compare the differenced series with the detrended series you calculated on A3. In other words, for the "Total Renewable Energy Production" compare the differenced series from Q1 with the series you detrended in Q2 using linear regression.

Using autoplot() + autolayer() create a plot that shows the three series together. 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.

```
#presenting three series
##Method1
plotorig <-autoplot(tsdata,series="Original")+ #this plot
  autolayer(renewable_detrend,series="Detrended")+ #no need to present this plot
  ylab("Renewable Energy") +
  ggtitle("Total Renewable Energy Production")
plotdif <-autoplot(diff_reneg, series="Differentiated")+ #and this plot
  autolayer(diff_reneg,series="Detrended")+
  ylab("Renewable_Energy") +
  ggtitle("Diff Renewable Energy Production")
plot_grid(plotorig, plotdif, nrow=2,ncol=1)
```




```
##Method2
autoplot(tsdata,series="Original")+
  autolayer(renewable_detrend,series="Detrended")+
  autolayer(diff_reneg, series="Difference")+
  ylab("Renewable Energy") +
  ggtitle("Combined Renewable Energy Production")
```

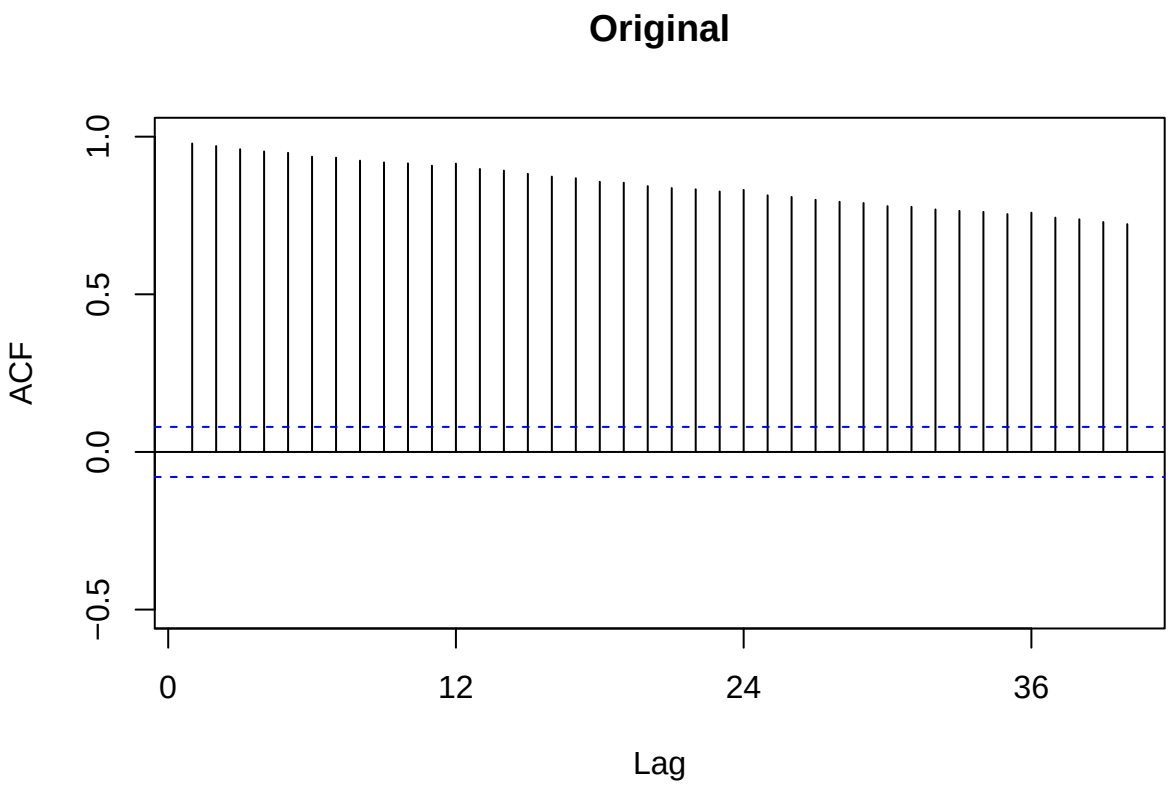


Q4

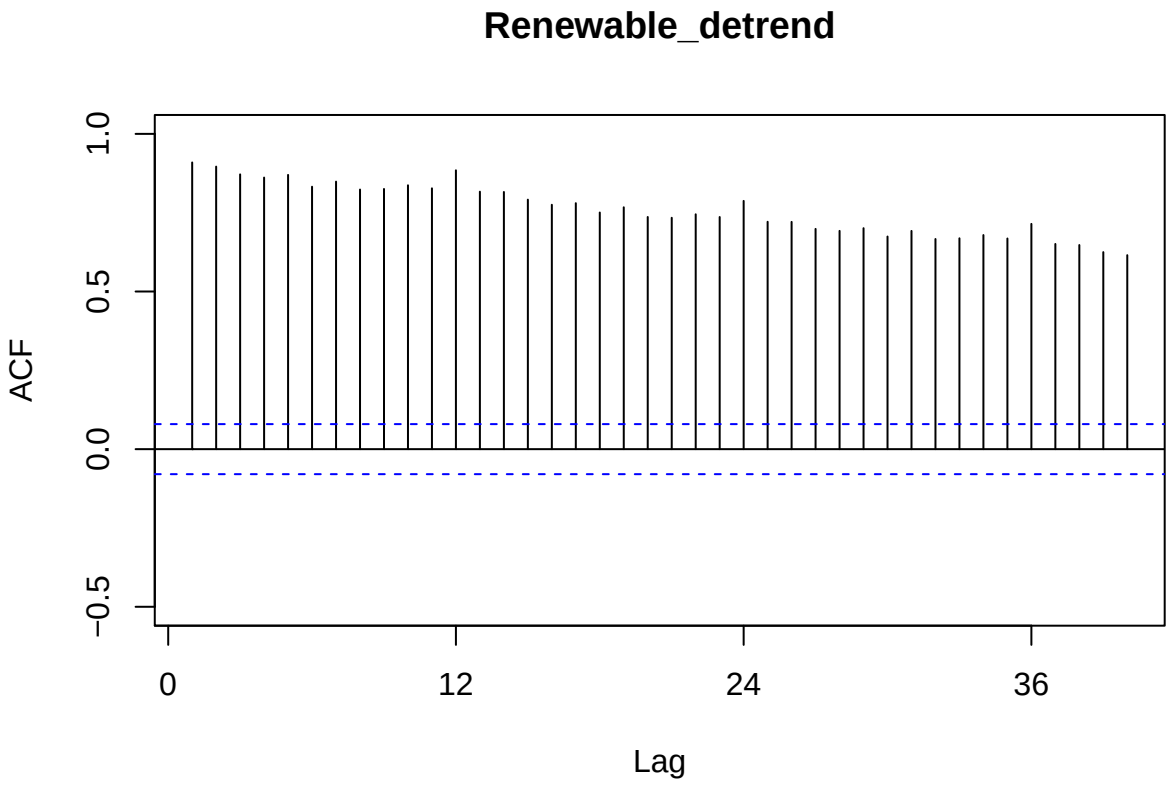
Plot the ACF for the three series and compare the plots. Add the argument `ylim=c(-0.5,1)` to the autoplot() or Acf() function - whichever you are using to generate the plots - to make sure all three y axis have the same limits. Which method do you think was more efficient in eliminating the trend? The linear regression or differencing?

```
plot_grid(
autoplot(Acf(tsdata,lag.max=40, ylim = c(-0.5, 1), main = "Original"), plot=FALSE),
autoplot(Acf(renewable_detrend, lag.max = 40, ylim = c(-0.5, 1), main = "Renewable_detrend"), plot=FALSE
autoplot(Acf(diff_reneg, lag.max = 40, ylim = c(-0.5, 1), main = "Renewable_difference"), plot=FALSE),
nrow=2,ncol=2
)
```

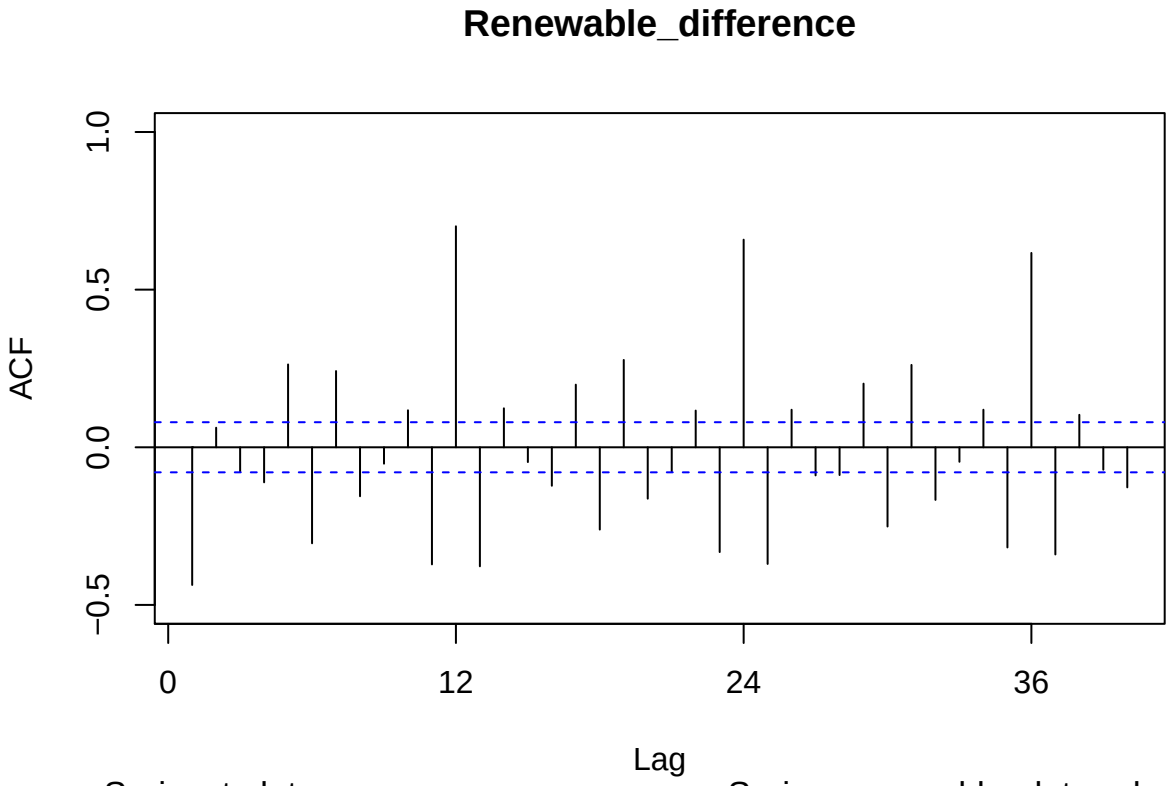
```
## Warning in ggplot2::geom_segment(lineend = "butt", ...): Ignoring unknown
## parameters: `plot`
```

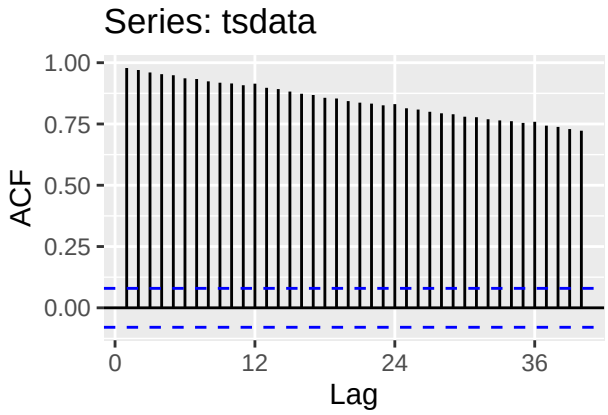


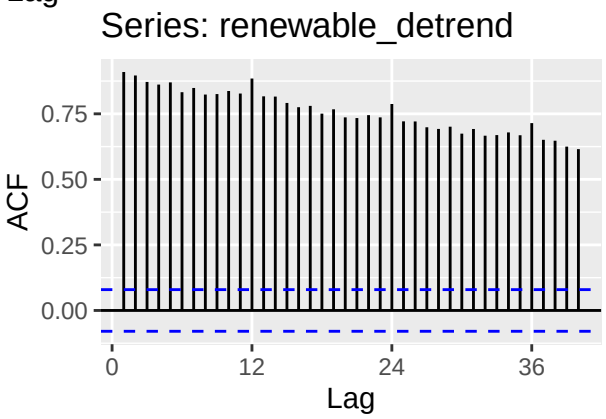

```
## Warning in ggplot2::geom_segment(lineend = "butt", ...): Ignoring unknown
## parameters: `plot`
```

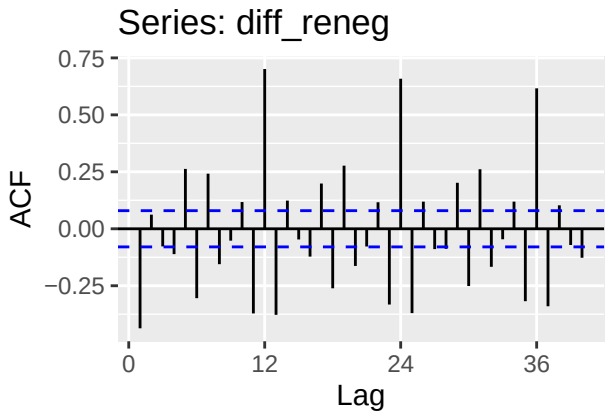



```
## Warning in ggplot2::geom_segment(lineend = "butt", ...): Ignoring unknown
## parameters: `plot`
```










```
#Explanation
# Acf shows the seasonal component very clearly.
```

Ploting the series, the difference method seems to detrend better than the regression method.

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.

```
#Seasonal Mann-Kendall test
smk_result <- SeasonalMannKendall(tsdata)
print("Results for Seasonal Mann Kendall")
```

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

```
print(summary(smk_result))
```

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

```
#ADF test
adf_result <- adf.test(tsdata,alternative = "stationary")
print("Results for ADF test")
```

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

```
print(adf_result)
```

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

Based on the Seasonal Mann-Kendall test, we can conclude that there is a statistically significant upward seasonal trend in our data (tau = 0.783, pvalue =< 2.22e-16). The Adf test's output (p-value = 0.9) is higher, therefore we cannot reject the HO of of a unit root (non-stationarity). Meaning there is likelihood of the presence of stochastic trend in our time series.

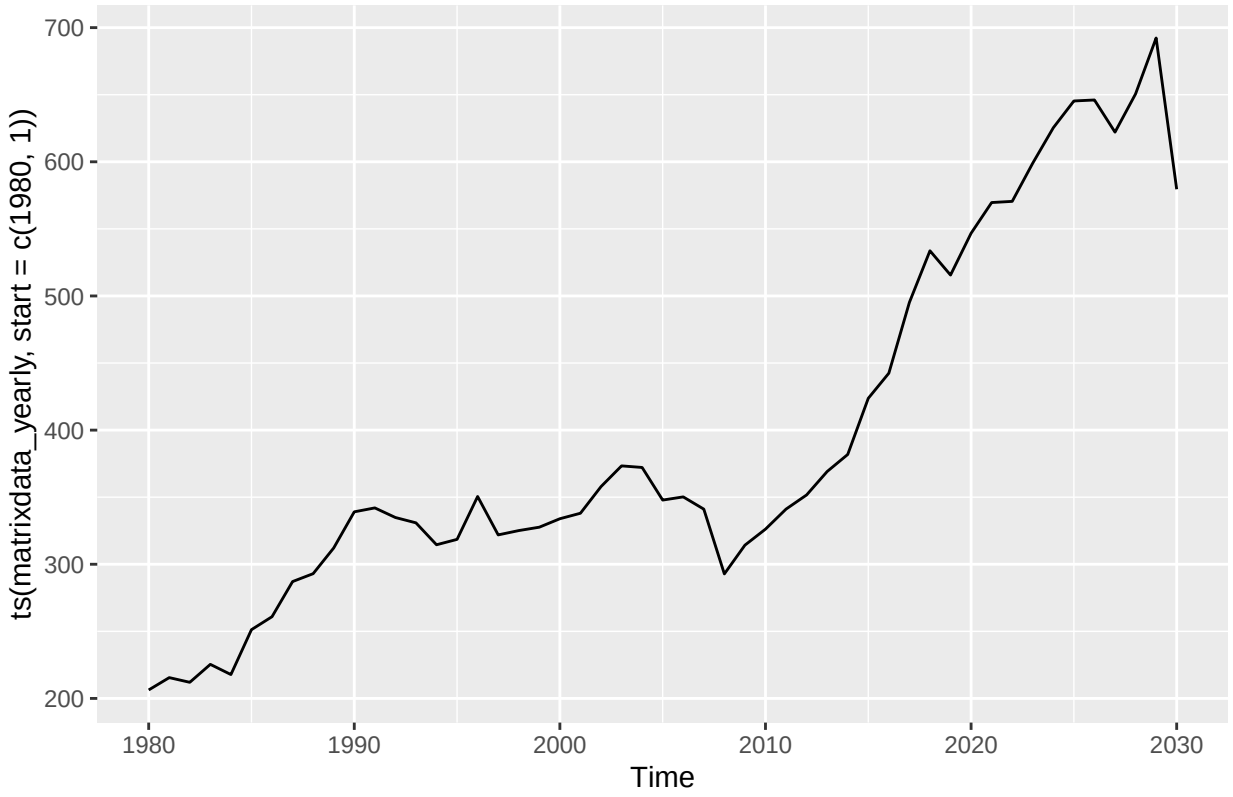
Q6

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().

```
# Aggregaring mean
matrixdata <- matrix(data$renen, byrow = FALSE, nrow = 12)
```

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

```
matrixdata_yearly <- colMeans(matrixdata)
autoplot(ts(matrixdata_yearly,start=c(1980,1)))
```



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?

```
#Seasonal Mann-Kendall test
smk_result <- SeasonalMannKendall(ts(matrixdata_yearly))
print("Results for 2nd Seasonal Mann Kendall")
```

```
## [1] "Results for 2nd Seasonal Mann Kendall"
```

```
print(summary(smk_result))
```

```
## Score =  1019 , Var(Score) = 15158.33
## denominator =  1275
## tau = 0.799, 2-sided pvalue =2.2204e-16
## NULL
```

```
#ADF test
adf_result <- adf.test(matrixdata_yearly,alternative = "stationary")
print("Results for ADF test")
```

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

```
print(adf_result)
```

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

This output also confirms our previous observation in Q6.