ENV 790.30 - Time Series Analysis for Energy Data | Spring 2021

Assignment 4 - Due date 02/25/21

Student Name

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 project open the first thing you will do is change "Student Name" on line 3 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. Rename the pdf file such that it includes your first and last name (e.g., "LuanaLima_TSA_A04_Sp21.Rmd"). Submit this pdf using Sakai.

Questions

Consider the same data you used for A2 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.

R packages needed for this assignment:"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(lubridate)
```

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

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

```
library(ggplot2)
library(forecast)
```

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

```
library(Kendall)
library(tseries)
library(outliers)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
## v tibble  3.0.6     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x lubridate::as.difftime() masks base::as.difftime()
## x lubridate::date()        masks base::date()
## x dplyr::filter()          masks stats::filter()
## x lubridate::intersect()   masks base::intersect()
## x dplyr::lag()             masks stats::lag()
## x lubridate::setdiff()     masks base::setdiff()
## x lubridate::union()       masks base::union()
```

```
library(dplyr)
```

Stochastic Trend and Stationarity Test

For this part you will once again work only with the following columns: Total Biomass Energy Production, Total Renewable Energy Production, Hydroelectric Power Consumption. Create a data frame structure with these three time series and the Date column. Don't forget to format the date object.

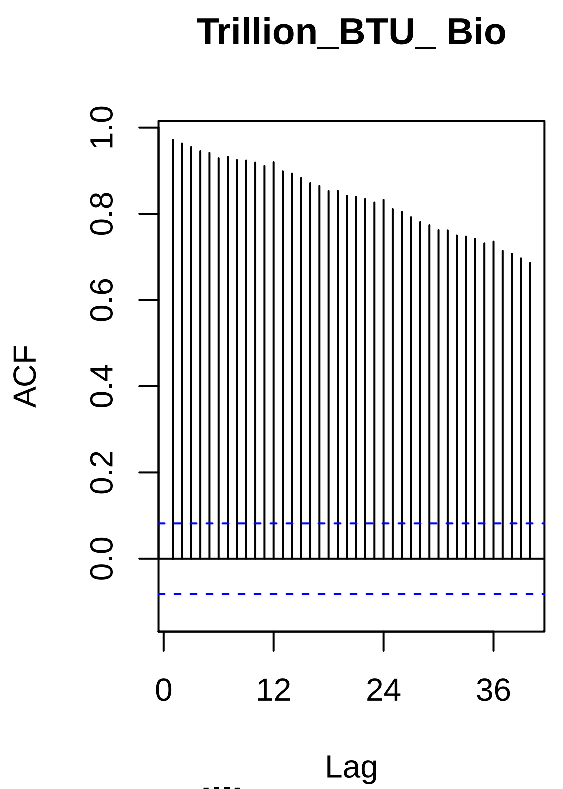
```
Table_1<- read.csv("../Data/table102.csv", header=TRUE)

first <-
  Table_1 %>%
  select("Month","Bio","Renew","Hydro") %>%
  mutate(Bio = as.numeric(Bio), Renew = as.numeric(Renew), Hydro = as.numeric(Hydro))%>%
  drop_na(Bio,Renew, Hydro)

my_date <- paste(first[,1])
my_date <- ym(my_date)

new_first <- cbind(my_date,first[,2:4])

ts_first <- ts(new_first[,2:4],frequency=12)
```

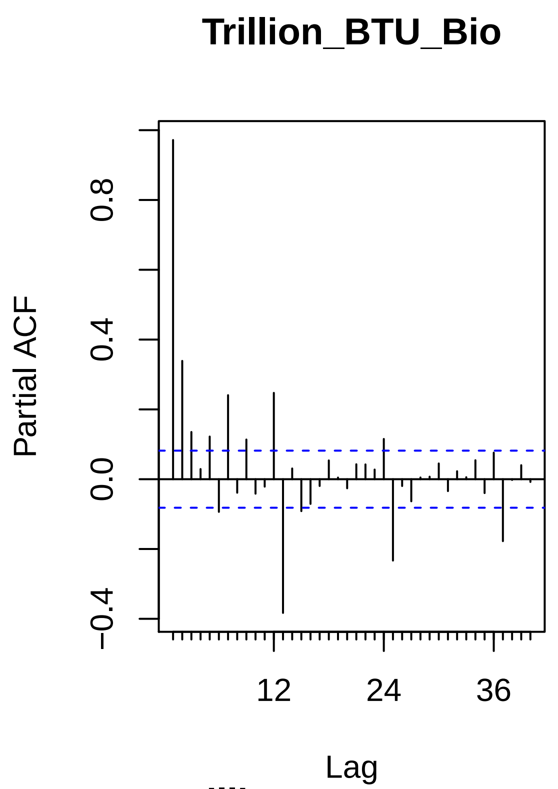
Lag1? ### Q1 Now let's try to difference these three series using function diff(). Start with the original data from part (b). Try differencing first at lag 1 and plot the remaining series. Did anything change? Do the series still seem to have trend?

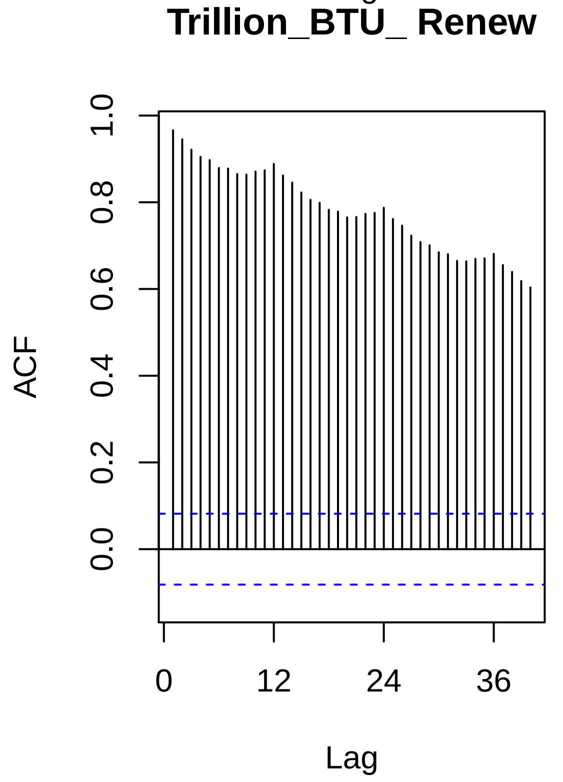
Yes the plot looks different than the orgional data series. The diff() removed the decreasing mean trend component in the biomas and renewable series as you know longer have highly correlated lags. However by removing the trend it appears that the seasonality component of the series is more pronounced. The Biomass and renewable energy production both appear to have some degree of seasonality.

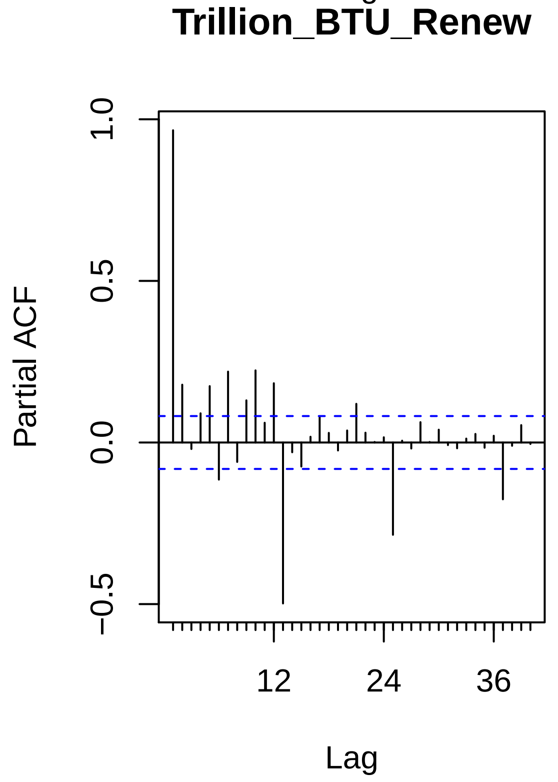
The hydro power production chart did not change very much if at all. The trend in this seris was not significant. The seasonality can be seen in both the diff and orgional plot)

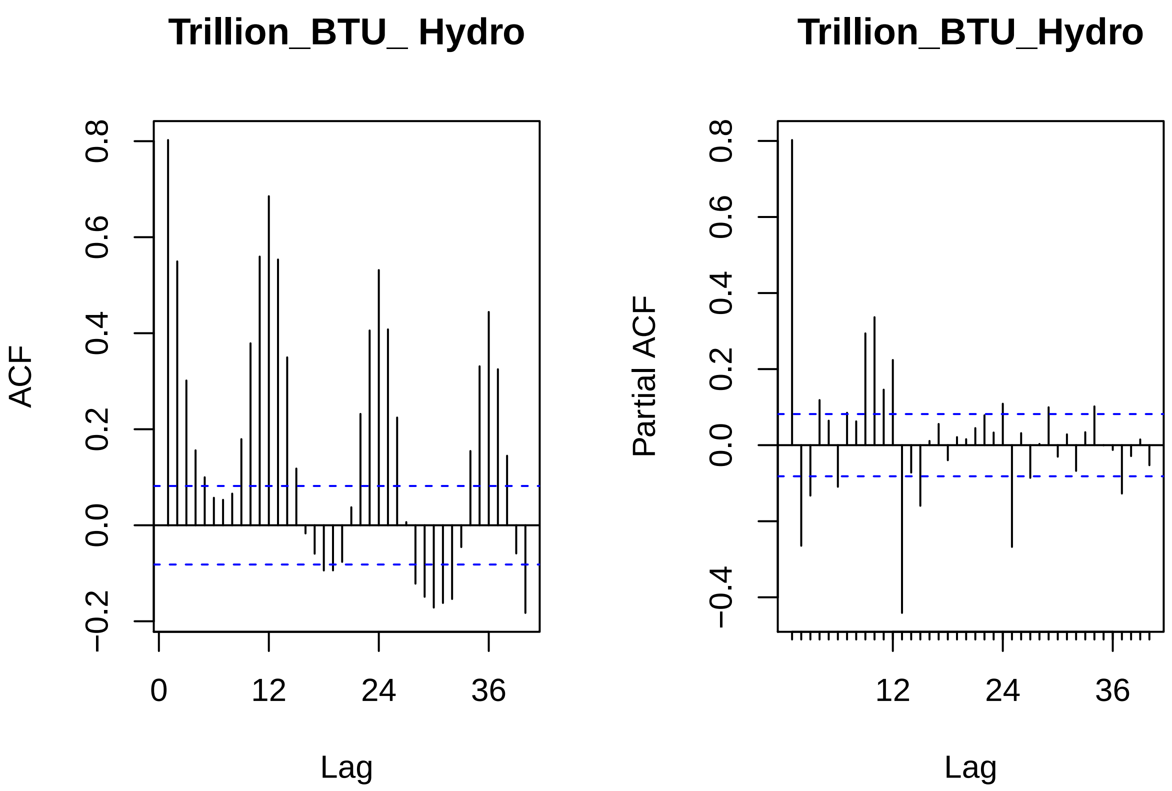
```
nenergy <- ncol(new_first)-1
for(i in 1:nenergy){
  par(mfrow=c(1,2))
  Acf(ts_first[,i],lag.max=40,main=paste("Trillion_BTU_ ",colnames(ts_first)[(i)],sep=""))
  Pacf(ts_first[,i],lag.max=40,main=paste("Trillion_BTU_",colnames(ts_first)[(i)],sep=""))}
```


Trillion_BTU_ Bio
ACF
1.0
0.8
0.6
0.4
0.2
0.0
0
12
24
36
Lag

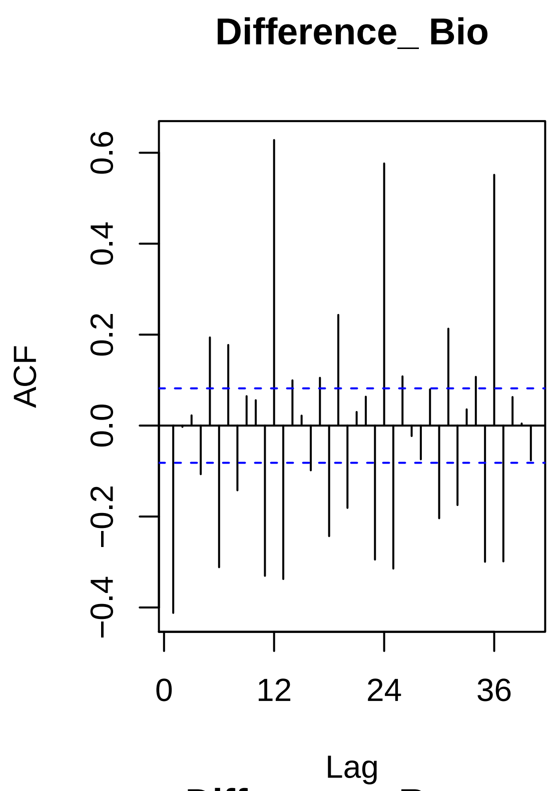

Trillion_BTU_Bio
Partial ACF
0.8
0.4
0.0
−0.4
12
24
36
Lag

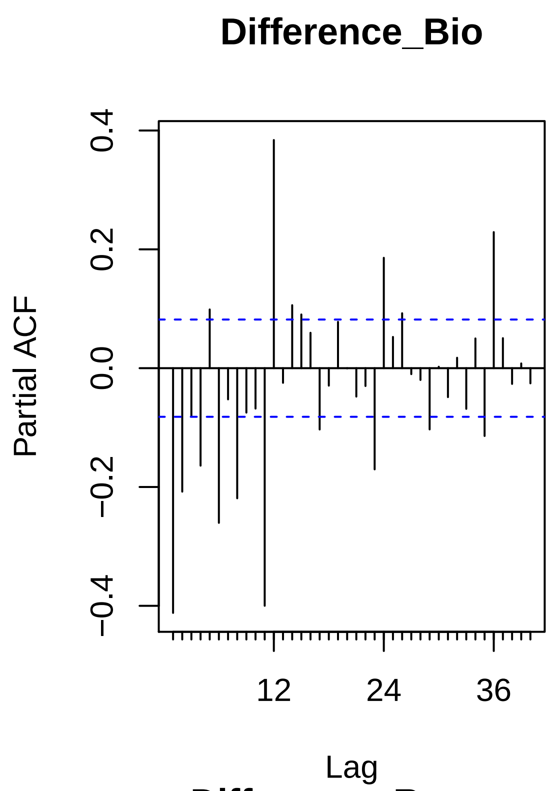

Trillion_BTU_ Renew
ACF
1.0
0.8
0.6
0.4
0.2
0.0
0
12
24
36
Lag

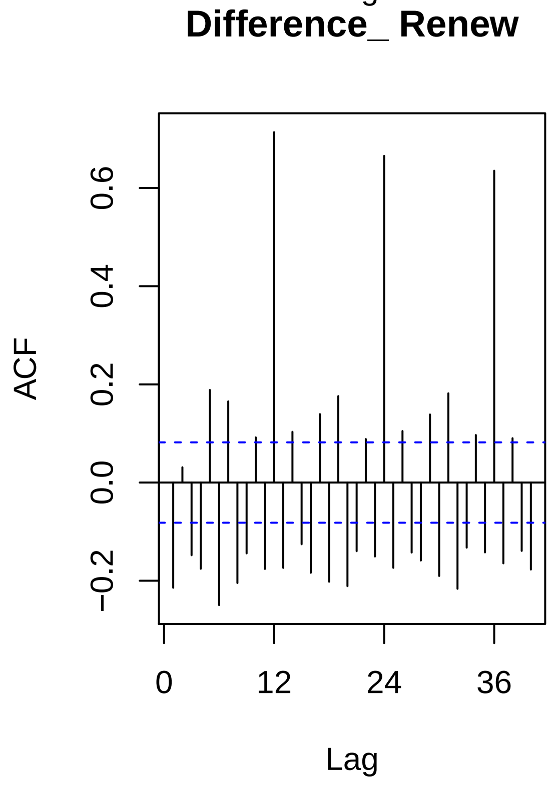

Trillion_BTU_Renew
Partial ACF
1.0
0.5
0.0
−0.5
12
24
36
Lag

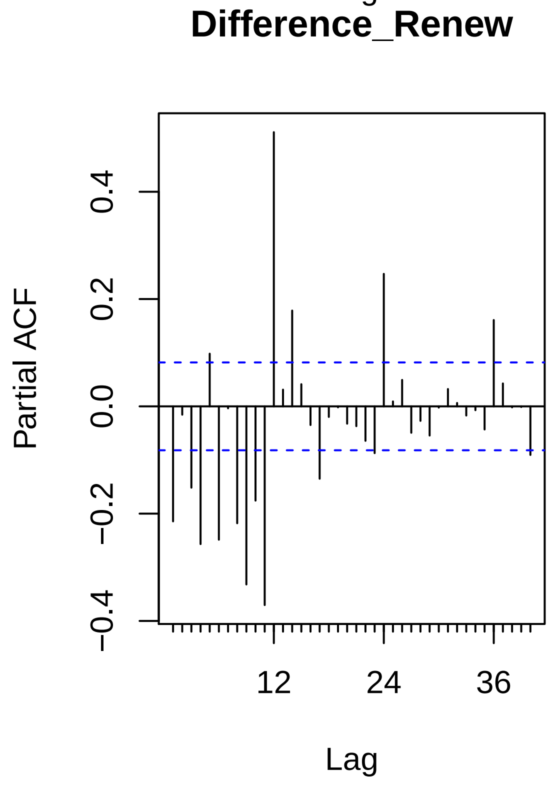


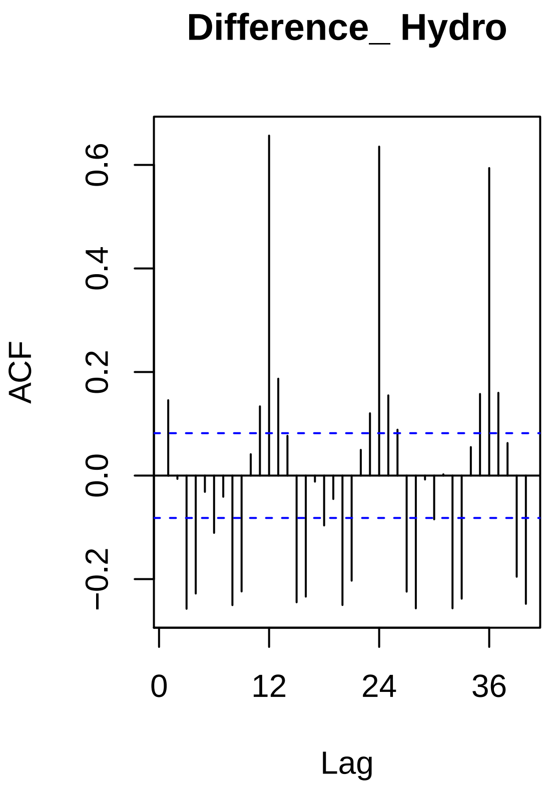
```r
difference_energy <- diff(ts_first, differences=1)

nenergy <- ncol(new_first)-1
nobs <- nrow(new_first)

for(i in 1:nenergy){
  par(mfrow=c(1,2))
  Acf(difference_energy[,i],lag.max=40,main=paste("Difference_ ",colnames(ts_first)[(i)],sep=""))
  Pacf(difference_energy[,i],lag.max=40,main=paste("Difference_",colnames(ts_first)[(i)],sep=""))
}
```

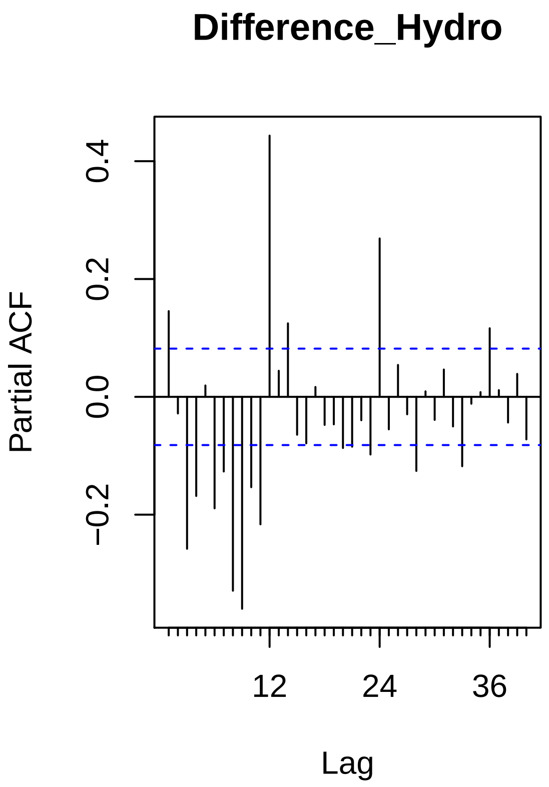

Difference_ Bio
ACF
Lag
0
12
24
36
−0.4
−0.2
0.0
0.2
0.4
0.6


Difference_Bio
Partial ACF
Lag
12
24
36
−0.4
−0.2
0.0
0.2
0.4


Difference_ Renew
ACF
Lag
0
12
24
36
−0.2
0.0
0.2
0.4
0.6


Difference_Renew
Partial ACF
Lag
12
24
36
−0.4
−0.2
0.0
0.2
0.4





I keep getting the sam seasonmannkendall value? ### Q2 Compute Mann-Kendall and Spearman's Correlation Rank Test for each time series. Ask R to print the results. Interpret the results.

Mann-kendall checks for deterministic trend . When running the test for each series the P value was less than 0.05 and I failed to reject the null, thus my series are not stationary, the data follows a trend for biomass, renewable and hydro production. The Score for the biomass energy was 116743 which is a high score, the score for renewable was 107415 which is also quite high, score indicates how many times the observation increased. The hydropower production score was negative indicating the the observation decrease over time.

Spearman's correlation Rank

The correlation coefficient in both test on the biomass series and renewable enrgy tend had a p value above 0.7, thus both correlations where significant. The p-value in both thest was less than 0.05 thus I failed to reject the null that the data is stationary, and both series have a trend.

The correlation coefficent was run on hydro power production. The P-valuewas -0.47 so while high it is not significant. The P value for this serries was also less than 0.05, thus I failed to reject the null, the hydro power production series is not stationary.

```r
for(i in 1:nenergy){
SMKtest <- MannKendall(ts_first[,i])
print(summary(SMKtest))
}
```

```
## Score =  116743 , Var(Score) = 21067998
## denominator =  164450
## tau = 0.71, 2-sided pvalue =< 2.22e-16
## NULL
## Score =  107415 , Var(Score) = 21068000
## denominator =  164451
## tau = 0.653, 2-sided pvalue =< 2.22e-16
## NULL
## Score =  -30759 , Var(Score) = 21068000
## denominator =  164451
## tau = -0.187, 2-sided pvalue =2.0685e-11
## NULL
```

```
# spearmna correlation rank test Bomass energy

bio_matrix <- matrix(ts_first[,1], byrow = FALSE, nrow=12)
```

```
## Warning in matrix(ts_first[, 1], byrow = FALSE, nrow = 12): data length [574] is
## not a sub-multiple or multiple of the number of rows [12]
```

```
bio_yearly <- colMeans(bio_matrix)

my_year=c(year(first(my_date)):year(last(my_date)))
new_bio_yearly <- data.frame(my_year, bio_yearly)

sp_rho=cor.test(bio_yearly,my_year,method = "spearman")
print(sp_rho)
```

```
##
##  Spearman's rank correlation rho
##
## data:  bio_yearly and my_year
## S = 2214, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8798307
```

```
#Renewable
renew_matrix <- matrix(ts_first[,2], byrow = FALSE, nrow=12)
```

```
## Warning in matrix(ts_first[, 2], byrow = FALSE, nrow = 12): data length [574] is
## not a sub-multiple or multiple of the number of rows [12]
```

```
renew_yearly <- colMeans(renew_matrix)
my_year=c(year(first(my_date)):year(last(my_date)))


sp_rho_2=cor.test(renew_yearly,my_year,method = "spearman")
print(sp_rho_2)
```

```
##
##  Spearman's rank correlation rho
##
## data:  renew_yearly and my_year
## S = 2560, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8610508
```

```
#Hydro
hydro_matrix <- matrix(ts_first[,3], byrow = FALSE, nrow=12)
```

```
## Warning in matrix(ts_first[, 3], byrow = FALSE, nrow = 12): data length [574] is
## not a sub-multiple or multiple of the number of rows [12]
```

```
hydro_yearly <- colMeans(hydro_matrix)
my_year=c(year(first(my_date)):year(last(my_date)))


sp_rho_3=cor.test(hydro_yearly,my_year,method = "spearman")
print(sp_rho_3)
```

```
##
##  Spearman's rank correlation rho
##
## data:  hydro_yearly and my_year
## S = 27250, p-value = 0.000661
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.4790491
```

Decomposing the series

For this part you will work only with the following columns: Solar Energy Consumption and Wind Energy Consumption.

Q3

unrecognizable date object when I create numeric and remove na?

Create a data frame structure with these two time series only and the Date column. Drop the rows with *Not Available* and convert the columns to numeric. You can use filtering to eliminate the initial rows or conver to numeric and then use the drop_na() function. If you are familiar with pipes for data wrangling, try using it!

```
Table_new1 <- read.csv("../Data/table102.csv", header=TRUE)

renewable <-
  Table_new1 %>%
  select("Month","Solar.Energy.Consumption","Wind.Energy.Consumption") %>%
  rename(Date = "Month", Solar_Consumption = "Solar.Energy.Consumption",Wind_Consumption = "Wind.Energy
  mutate(Solar_Consumption = as.numeric(Solar_Consumption), Wind_Consumption = as.numeric(Wind_Consumpti
```

```
## Warning: Problem with `mutate()` input `Solar_Consumption`.
## i NAs introduced by coercion
## i Input `Solar_Consumption` is `as.numeric(Solar_Consumption)`.

## Warning: Problem with `mutate()` input `Wind_Consumption`.
## i NAs introduced by coercion
## i Input `Wind_Consumption` is `as.numeric(Wind_Consumption)`.
```

```
my_date <- paste(renewable[,1])
my_date <- ym(my_date)

new_renewable <- cbind(my_date,renewable[,2:3])

ts_renewable <- ts(new_renewable[,2:3],frequency=12)

ncolumn <- ncol(new_renewable)-1
nobs <- nrow(new_renewable)
```

Q4

Need to plot graph without nas?

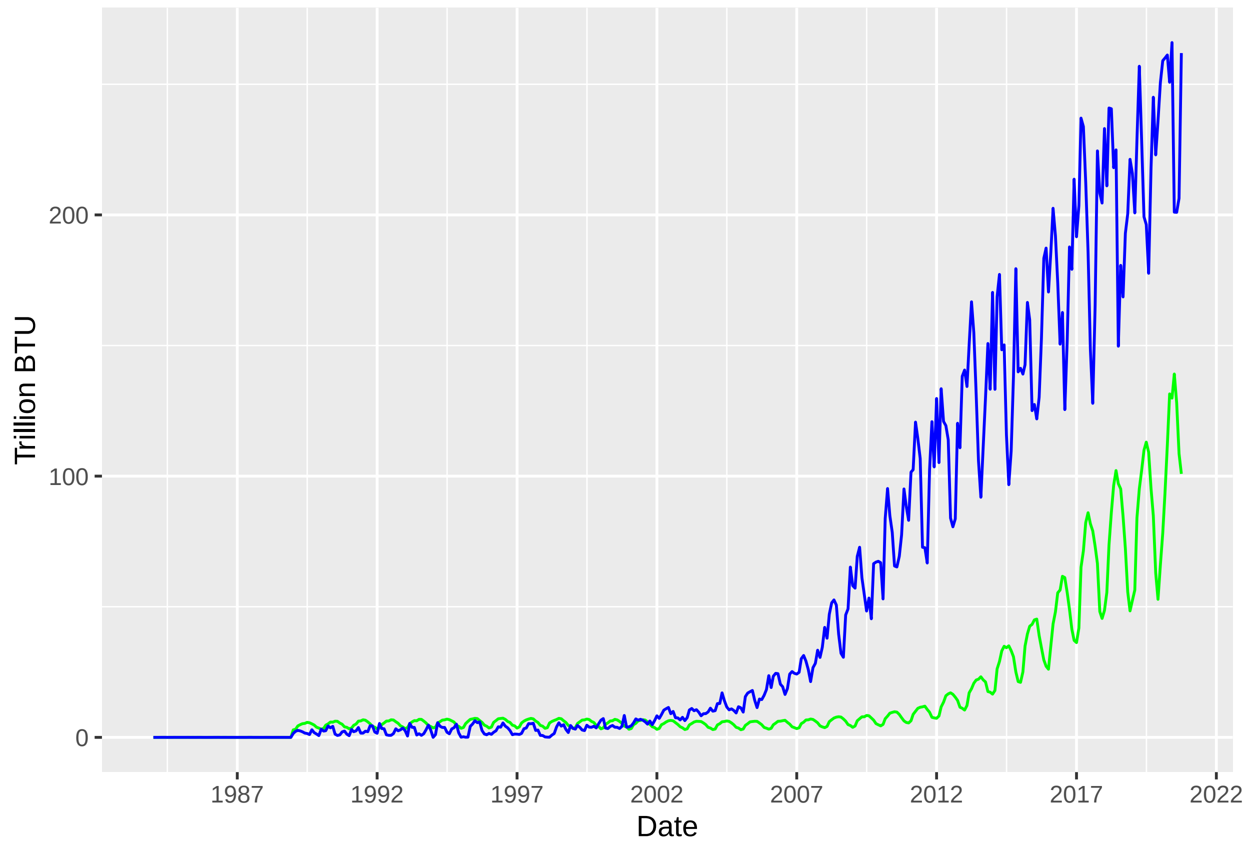
Plot the Solar and Wind energy consumption over time using ggplot. Explore the function scale_x_date() on ggplot and see if you can change the x axis to improve your plot. Hint: use *scale_x_date(date_breaks = "5 years", date_labels = "%Y")")*

Try changing the color of the wind series to blue. Hint: use *color = "blue"*

```
ggplot(data=new_renewable, aes(x=my_date, y=new_renewable$Solar_Consumption)) + geom_line(color="green")
            ylab(paste0("Trillion BTU")) +xlab(paste0("Date"))+ geom_line(aes(y=new_renewable$Wind_Consu
```

```
## Warning: Use of `new_renewable$Solar_Consumption` is discouraged. Use
## `Solar_Consumption` instead.
```

```
## Warning: Use of `new_renewable$Wind_Consumption` is discouraged. Use
## `Wind_Consumption` instead.
```



Q5

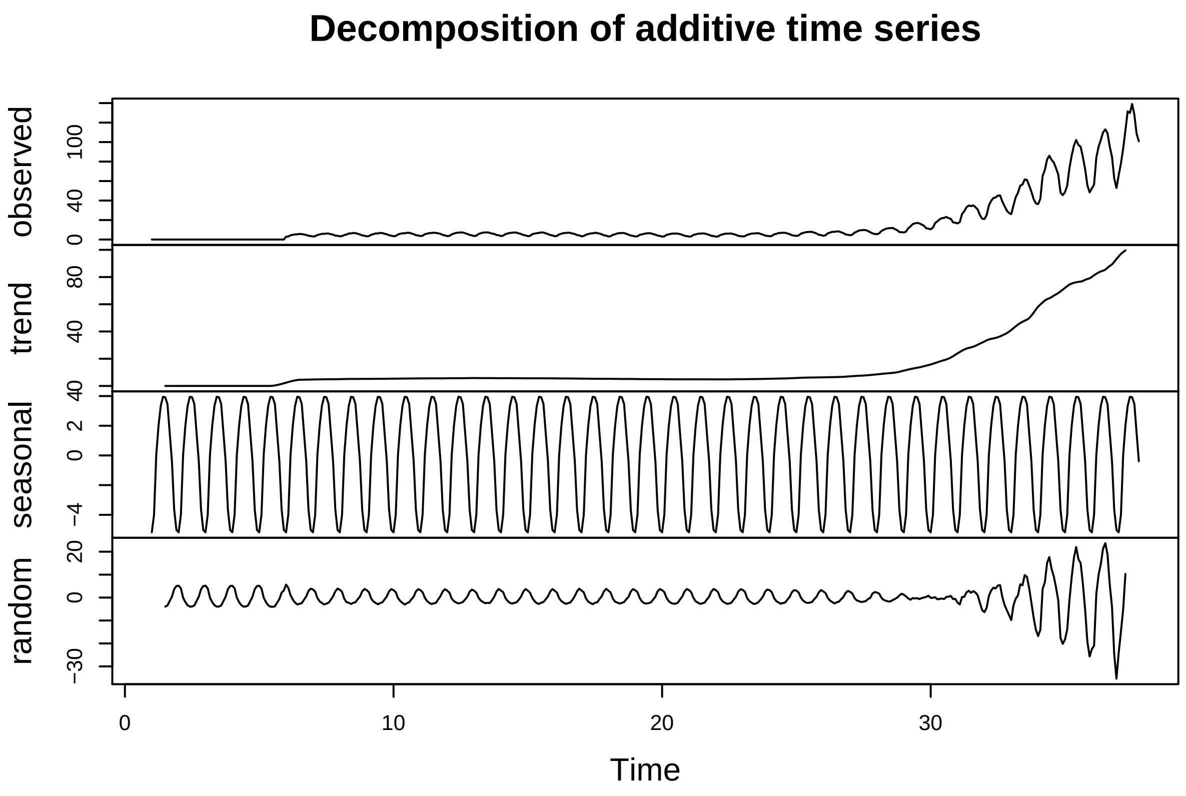
Transform wind and solar series into a time series object and apply the decompose function on them using the additive option. What can you say about the trend component? What about the random component? Does the random component look random? Or does it appear to still have some seasonality on it?

The trend component for the increases in the last half to last quater of the time series for both wind and solar. There is very little increase across the means in the first half of the plot but then there is a uptick in the means for the later times.

For the solar component the random seems to still have a seasonality component to it as it varies from positive to negative in a very repetitive manner. When you reach time 30 the random component has a greater absolute variance but still seems to have a seasonal trend. While the first observations fluxuate seasonally in the wind component once you reach 25 the wind random component looks more naturally random and does not appear to have a clear seasonality trend.

```r
#Solar
i=1
decompose_solar=decompose(ts_renewable[,i],"additive")
solar_random <- decompose_solar$random
plot(decompose_solar, x.y.labels= "Solar_Production")
```

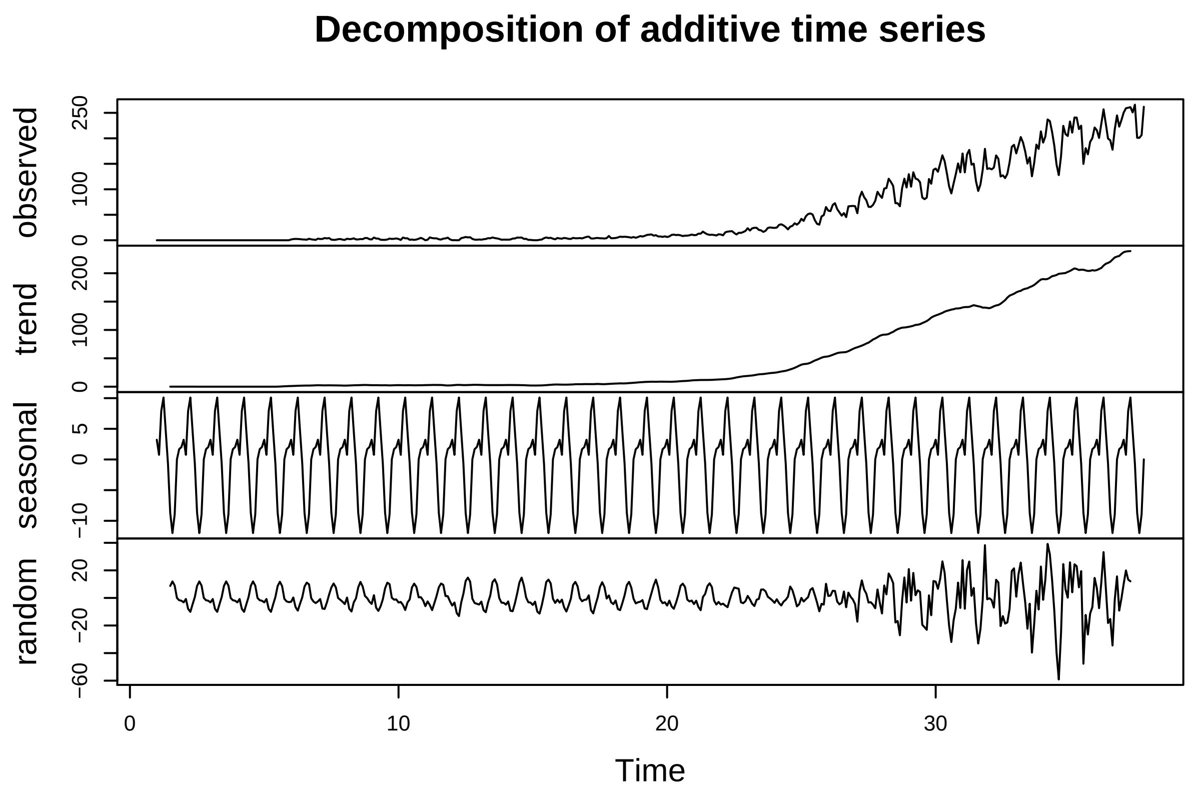
```
## Warning in plot.window(...): "x.y.labels" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "x.y.labels" is not a graphical parameter

## Warning in title(...): "x.y.labels" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "x.y.labels" is not a
## graphical parameter

## Warning in box(...): "x.y.labels" is not a graphical parameter

## Warning in axis(y.side, xpd = NA, cex.axis = cex.axis, col.axis = col.axis, :
## "x.y.labels" is not a graphical parameter

## Warning in mtext(nm[i], y.side, line = 3, cex = cex.lab, col = col.lab, :
## "x.y.labels" is not a graphical parameter

## Warning in plot.window(...): "x.y.labels" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "x.y.labels" is not a graphical parameter

## Warning in title(...): "x.y.labels" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "x.y.labels" is not a
## graphical parameter

## Warning in box(...): "x.y.labels" is not a graphical parameter

## Warning in axis(y.side, xpd = NA, cex.axis = cex.axis, col.axis = col.axis, :
## "x.y.labels" is not a graphical parameter

## Warning in mtext(nm[i], y.side, line = 3, cex = cex.lab, col = col.lab, :
## "x.y.labels" is not a graphical parameter

## Warning in plot.window(...): "x.y.labels" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "x.y.labels" is not a graphical parameter

## Warning in title(...): "x.y.labels" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "x.y.labels" is not a
## graphical parameter

## Warning in box(...): "x.y.labels" is not a graphical parameter

## Warning in axis(y.side, xpd = NA, cex.axis = cex.axis, col.axis = col.axis, :
## "x.y.labels" is not a graphical parameter

## Warning in mtext(nm[i], y.side, line = 3, cex = cex.lab, col = col.lab, :
## "x.y.labels" is not a graphical parameter

## Warning in plot.window(...): "x.y.labels" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "x.y.labels" is not a graphical parameter

## Warning in title(...): "x.y.labels" is not a graphical parameter
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "x.y.labels" is not a
## graphical parameter

## Warning in box(...): "x.y.labels" is not a graphical parameter

## Warning in axis(y.side, xpd = NA, cex.axis = cex.axis, col.axis = col.axis, :
## "x.y.labels" is not a graphical parameter

## Warning in axis(1, xpd = NA, cex.axis = cex.axis, col.axis = col.axis, font.axis
## = font.axis, : "x.y.labels" is not a graphical parameter

## Warning in mtext(nm[i], y.side, line = 3, cex = cex.lab, col = col.lab, :
## "x.y.labels" is not a graphical parameter

## Warning in mtext(xlab, side = 1, line = 3, cex = cex.lab, col = col.lab, :
## "x.y.labels" is not a graphical parameter

## Warning in mtext(main, side = 3, line = 3, cex = cex.main, font = font.main, :
## "x.y.labels" is not a graphical parameter
```



```
mean_solar <- mean(na.exclude(solar_random))
sd_solar <- sd(na.exclude(solar_random))
cat(mean_solar,sd_solar)
```

```
## -0.1269923 5.60796
```

```
#Using R decompose function for wind
i=2
decompose_wind=decompose(ts_renewable[,i],"additive")
plot(decompose_wind)
```



```
#Inspect random component
wind_random <- decompose_wind$random
mean_wind <- mean(na.exclude(wind_random))
sd_wind <- sd(na.exclude(wind_random))
cat(mean_wind,sd_wind)
```

```
## 0.08327103 11.04757
```

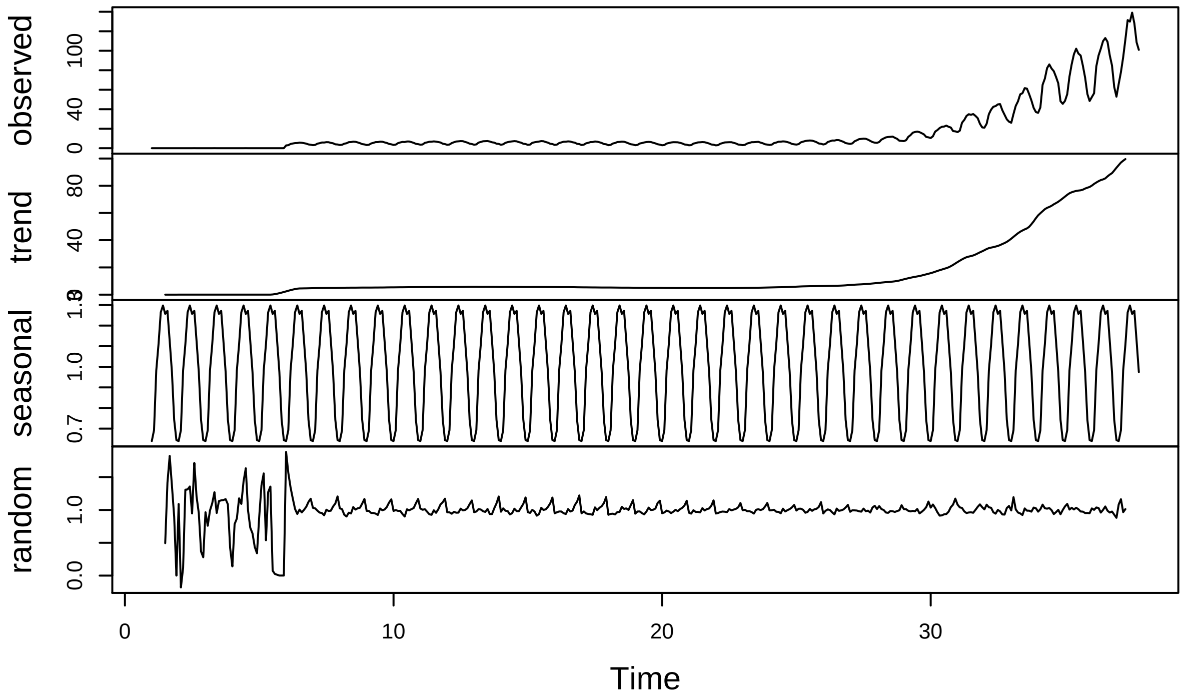
Q6

Use the decompose function again but now change the type of the seasonal component from additive to multiplicative. What happened to the random component this time?

The random component under the multiplicative seems to flip. Meaning in that the randomness is apparent in solar production random component for the first time readings 0-5, how after this the series seems to vary seasonally and follows the same low variance throughout the remaining time readings. In the wind production random component, the variance appears random from time readings 0-18. After this the variance is significantly less in terms of absolute variance. However there does not to be a clear seasonal trend but certainly appears less random.

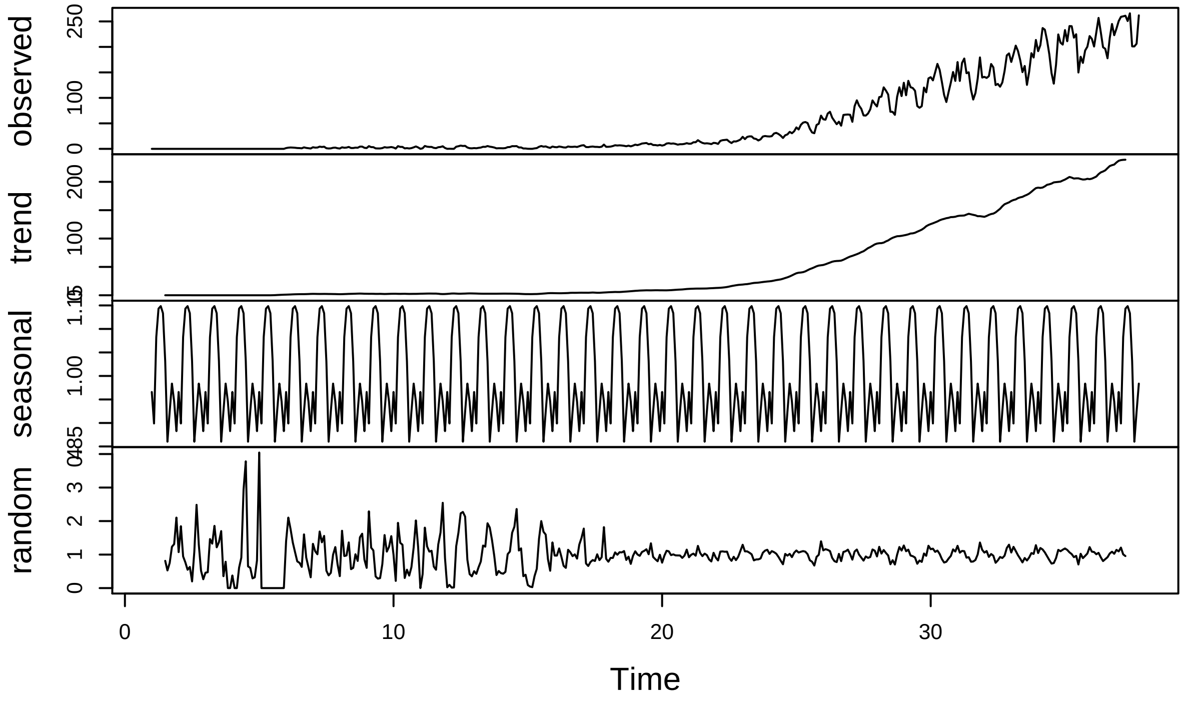
```
#Solar
i=1
decompose_solar_m=decompose(ts_renewable[,i],"multiplicative")
plot(decompose_solar_m)
```

Decomposition of multiplicative time series



```
#Wind
i=2
decompose_wind_m=decompose(ts_renewable[,i],"multiplicative")
plot(decompose_wind_m)
```

Decomposition of multiplicative time series



Q7

When fitting a model to this data, do you think you need all the historical data? Think about the date from 90s and early 20s. Are there any information from those year we might need to forecast the next six months of Solar and/or Wind consumption. Explain your response.

I think we do not need all of the historical data. The random component occurs later in the time serries and would likely be more valuable to forecast data from early 20s. The decompose function which was what was performed above utilizes a moving average technique, the decompose function specifies the frequency is 12, finding the average of 6 prior and 6 next observations of a specific time. The decompose is finding the 6 prior observations to compute the mean. That is why you get na for the fist and last 6 of the data set. The moving average replaces observations, with the mean. Thus it will be important to forcast last 6 of the series as data set increases with time and it might not be the best to simply replace the last 6 with the mean.