Assignment 7: Time Series Analysis

Amrita Sood

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on time series analysis.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Fay_A07_TimeSeries.Rmd") prior to submission.

The completed exercise is due on Tuesday, March 16 at 11:59 pm.

Set up

1. Set up your session:

- Check your working directory
- Load the tidyverse, lubridate, zoo, and trend packages
- Set your ggplot theme

2. Import the ten datasets from the Ozone_TimeSeries folder in the Raw data folder. These contain ozone concentrations at Garinger High School in North Carolina from 2010-2019 (the EPA air database only allows downloads for one year at a time). Import these either individually or in bulk and then combine them into a single dataframe named `GaringerOzone` of 3589 observation and 20 variables.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.5     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

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

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

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

```r
#install.packages("trend")
library(trend)
#install.packages("zoo")
library(zoo)
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
#install.packages("Kendall")
library(Kendall)
#install.packages("tseries")
library(tseries)
```

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

```r
getwd()
```

```
## [1] "/Users/amritasood/Desktop/CLASSES MEM/Spring 2021/EDA - ENV872/Environmental_Data_Analytics_202
```

```r
#1
EPA_2010 <-
  read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2010_raw.csv", stringsAsFactors = TRUE)
EPA_2011 <-
  read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2011_raw.csv", stringsAsFactors = TRUE)
EPA_2012 <-
  read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2012_raw.csv", stringsAsFactors = TRUE)
EPA_2013 <-
  read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2013_raw.csv", stringsAsFactors = TRUE)
EPA_2014 <-
  read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2014_raw.csv", stringsAsFactors = TRUE)
EPA_2015 <-
  read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2015_raw.csv", stringsAsFactors = TRUE)
EPA_2016 <-
  read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2016_raw.csv", stringsAsFactors = TRUE)
EPA_2017 <-
  read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2017_raw.csv", stringsAsFactors = TRUE)
EPA_2018 <-
  read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2018_raw.csv", stringsAsFactors = TRUE)
EPA_2019 <-
  read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2019_raw.csv", stringsAsFactors = TRUE)
GaringerOzone <- rbind(EPA_2010, EPA_2011, EPA_2012, EPA_2013, EPA_2014,EPA_2015, EPA_2016, EPA_2017,EPA
```

Wrangle

3. Set your date column as a date class.

4. Wrangle your dataset so that it only contains the columns Date, Daily.Max.8.hour.Ozone.Concentration, and DAILY_AQI_VALUE.

5. Notice there are a few days in each year that are missing ozone concentrations. We want to generate a daily dataset, so we will need to fill in any missing days with NA. Create a new data frame that

contains a sequence of dates from 2010-01-01 to 2019-12-31 (hint: `as.data.frame(seq())`). Call this new data frame Days. Rename the column name in Days to "Date".

6. Use a `left_join` to combine the data frames. Specify the correct order of data frames within this function so that the final dimensions are 3652 rows and 3 columns. Call your combined data frame GaringerOzone.

```
# 3
GaringerOzone$Date<-as.Date(GaringerOzone$Date, format = "%m/%d/%Y")



# 4

GaringerOzone <- GaringerOzone %>%
  select(Date, Daily.Max.8.hour.Ozone.Concentration, DAILY_AQI_VALUE)



# 5
Days<- GaringerOzone %>%
  mutate(Date=as.Date(Date))%>%
  complete(Date=seq.Date(min(Date), max(Date),by= "day"))




# 6
GaringerOzone <- left_join(Days, GaringerOzone)
```

```
## Joining, by = c("Date", "Daily.Max.8.hour.Ozone.Concentration", "DAILY_AQI_VALUE")
```

Visualize

7. Create a line plot depicting ozone concentrations over time. In this case, we will plot actual concentrations in ppm, not AQI values. Format your axes accordingly. Add a smoothed line showing any linear trend of your data. Does your plot suggest a trend in ozone concentration over time?

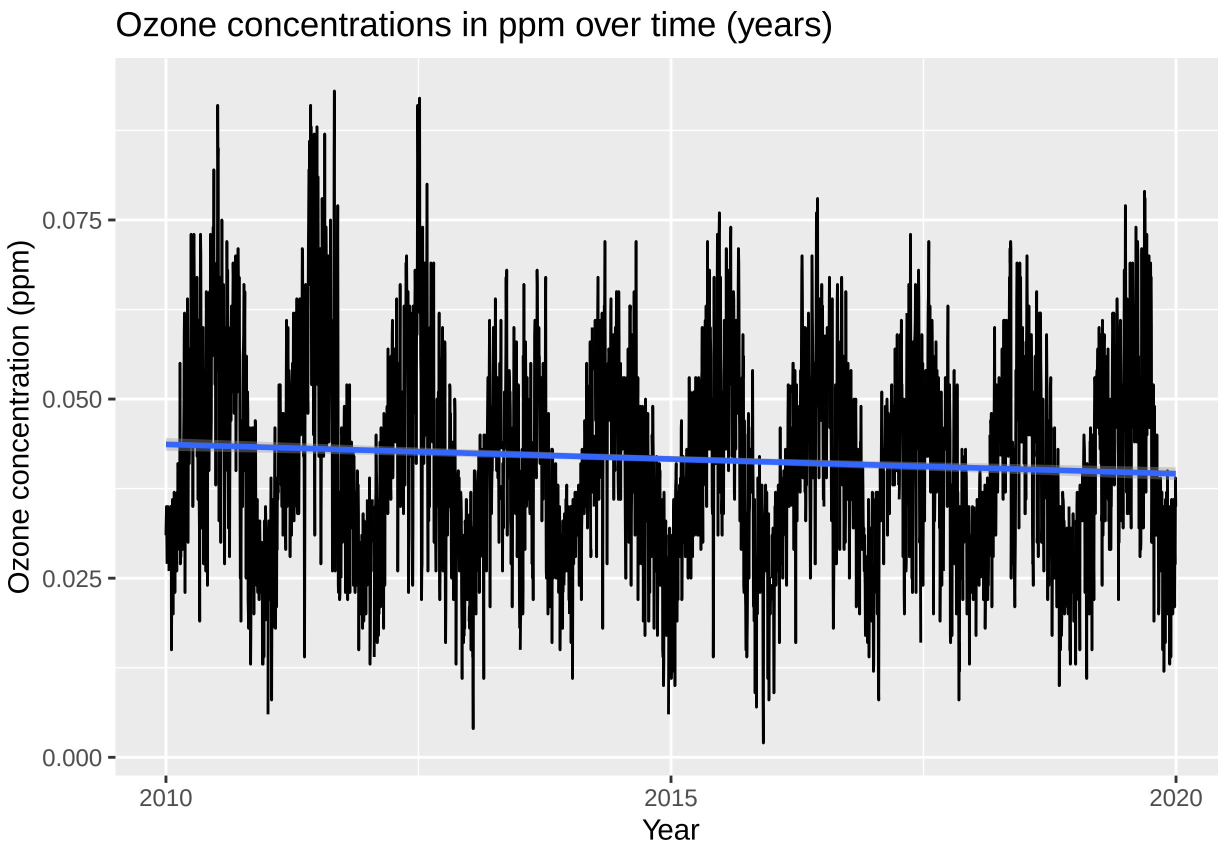
```
#7

Plot1 <- ggplot(GaringerOzone,
                     aes(x = Date, y = Daily.Max.8.hour.Ozone.Concentration)) +
  geom_line() +
  labs(title = "Ozone concentrations in ppm over time (years)",
       y = "Ozone concentration (ppm)", x = "Year") +
  geom_smooth(method = lm)


  print(Plot1)
```

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

```
## Warning: Removed 63 rows containing non-finite values (stat_smooth).
```



Answer: Yes, the plot suggests seasonality

Time Series Analysis

Study question: Have ozone concentrations changed over the 2010s at this station?

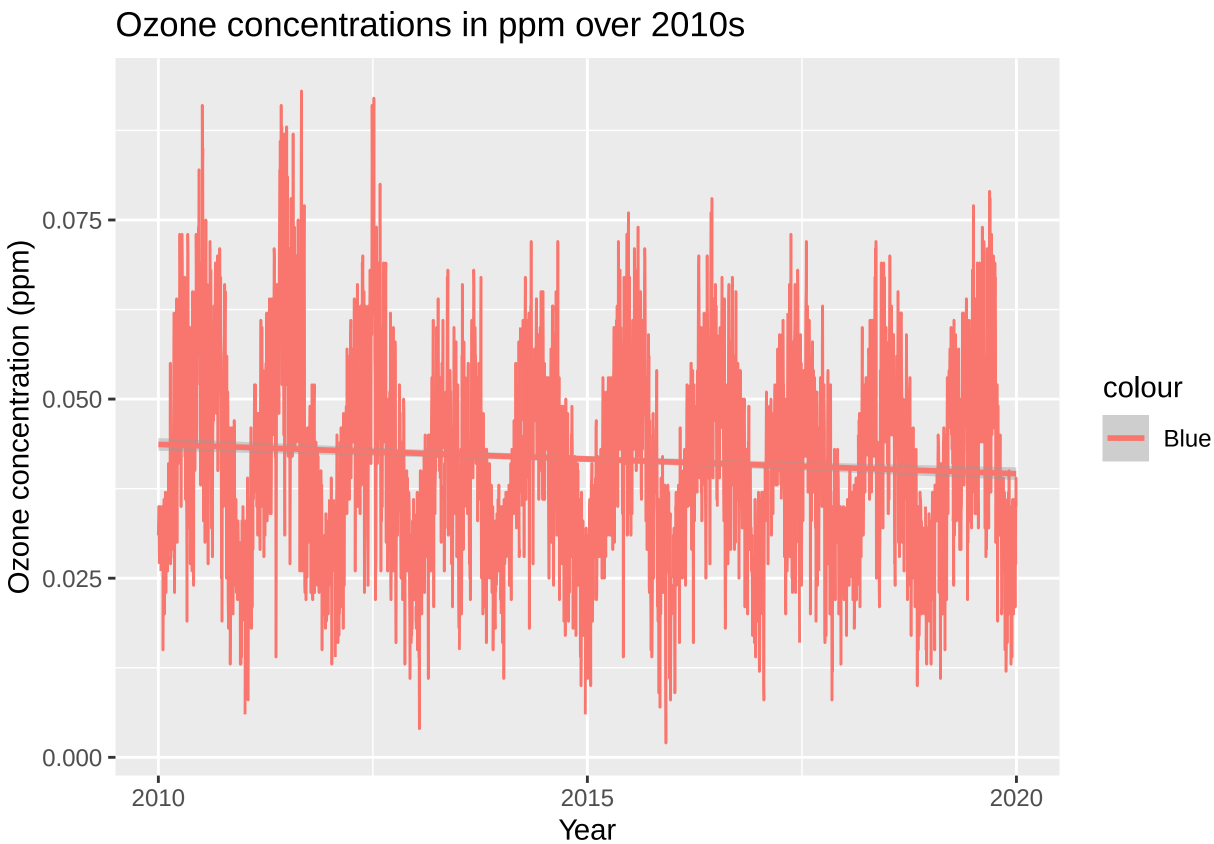
8. Use a linear interpolation to fill in missing daily data for ozone concentration. Why didn't we use a piecewise constant or spline interpolation?

```
#8
GaringerOzone.clean <-
  GaringerOzone %>%
  mutate( Concentrations = zoo::na.approx(Daily.Max.8.hour.Ozone.Concentration) )
Plot2 <- ggplot(GaringerOzone,
                     aes(x = Date, y = Daily.Max.8.hour.Ozone.Concentration, color= "Blue")) +
  geom_line() +
  labs(title = "Ozone concentrations in ppm over 2010s",
       y = "Ozone concentration (ppm)", x = "Year") +
  geom_smooth(method = lm)


  print(Plot2)
```

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

```
## Warning: Removed 63 rows containing non-finite values (stat_smooth).
```



Answer: Yes, ozone concetration has changed over the years. Piecewise constant is used when the missing data are assumed to be equal and spline is when a quadratic function is used to interpolate rather than a straight line.Linear is most appropriate to sort of draw a line between the missing data points.

9. Create a new data frame called `GaringerOzone.monthly` that contains aggregated data: mean ozone concentrations for each month. In your pipe, you will need to first add columns for year and month to form the groupings. In a separate line of code, create a new Date column with each month-year combination being set as the first day of the month (this is for graphing purposes only)

```
#9
GaringerOzone.monthly<-GaringerOzone.clean %>%
  mutate(Year = year(Date)) %>%
  mutate(Month = month(Date)) %>%
  group_by(Month, Year) %>%
  summarise(MeanConcentrations = mean(Concentrations)) %>%
  arrange(Year, Month)
```

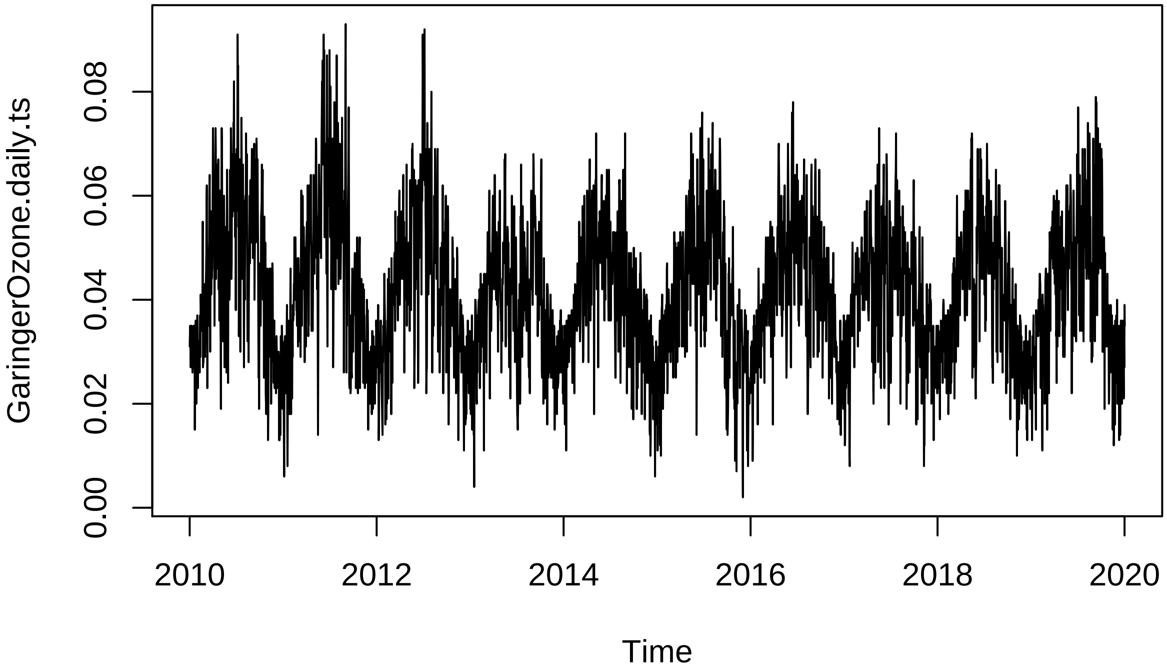
```
## `summarise()` has grouped output by 'Month'. You can override using the `.groups` argument.
```

```
GaringerOzone.monthly <- GaringerOzone.monthly %>%
  mutate(Date = my(paste(Month,"-",Year)))
```

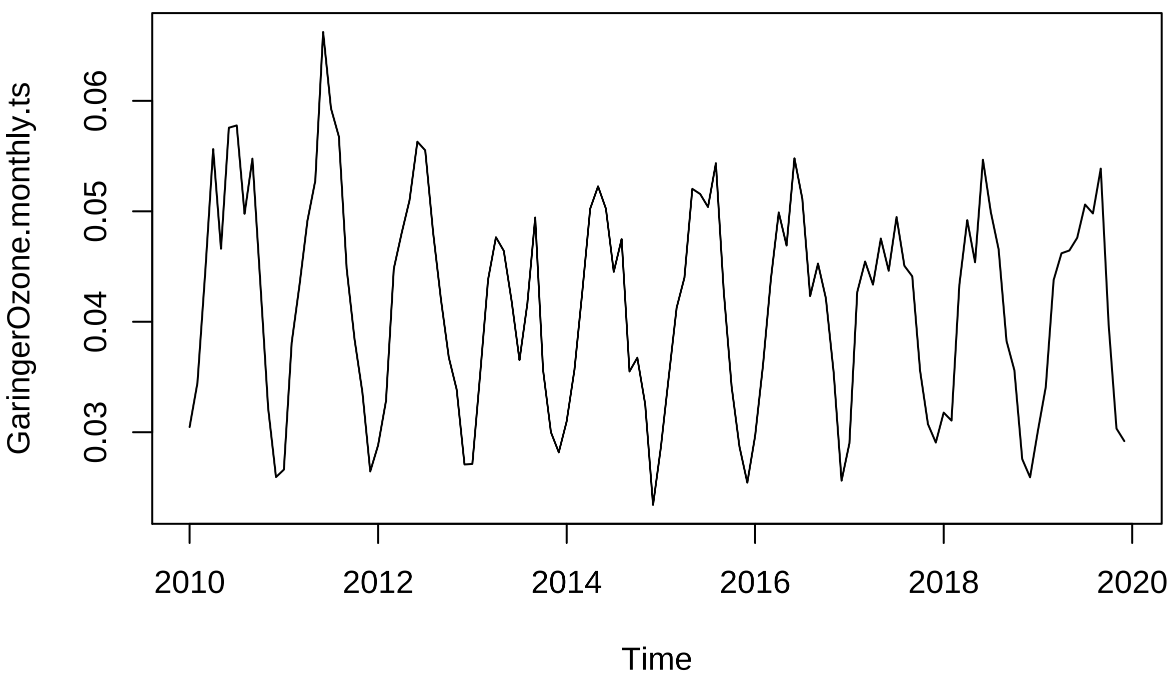
10. Generate two time series objects. Name the first `GaringerOzone.daily.ts` and base it on the dataframe of daily observations. Name the second `GaringerOzone.monthly.ts` and base it on the monthly average ozone values. Be sure that each specifies the correct start and end dates and the frequency of the time series.

```
#10
GaringerOzone.daily.ts <-  ts(GaringerOzone.clean$Concentrations,
```

```
                                  start = c(2010,1), frequency = 365)
plot(GaringerOzone.daily.ts)
```



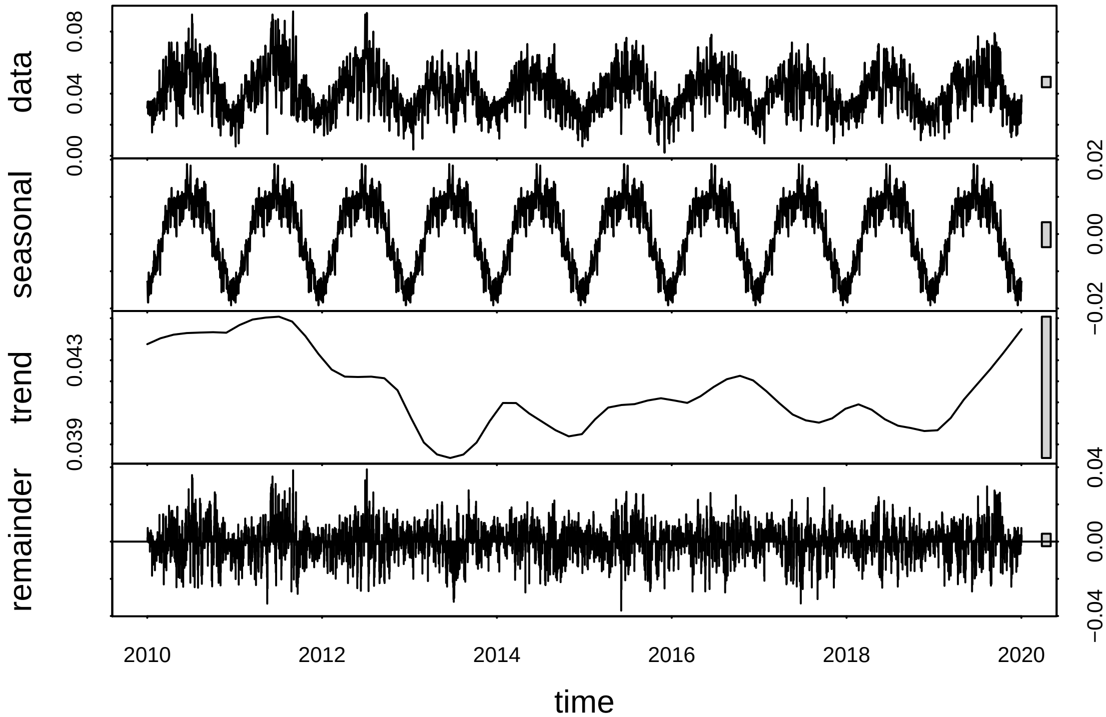

```
GaringerOzone.monthly.ts <- ts(GaringerOzone.monthly$MeanConcentrations,
                                  start = c(2010,1), frequency = 12)
plot(GaringerOzone.monthly.ts)
```



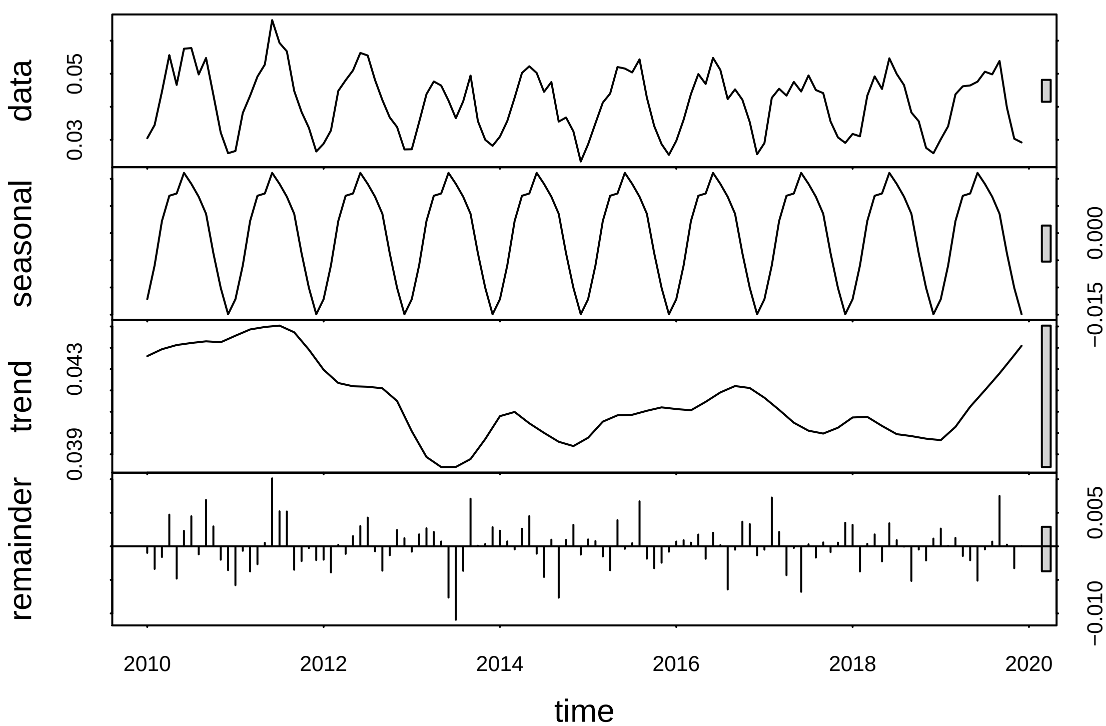
11. Decompose the daily and the monthly time series objects and plot the components using the `plot()` function.

```
#11
Decomposed_daily <- stl(GaringerOzone.daily.ts, s.window = "periodic")
```

```
plot(Decomposed_daily)
```




```
Decomposed_monthly <- stl(GaringerOzone.monthly.ts, s.window = "periodic")
plot(Decomposed_monthly)
```



12. Run a monotonic trend analysis for the monthly Ozone series. In this case the seasonal Mann-Kendall is most appropriate; why is this?

```r
#12

trend1 <- Kendall::SeasonalMannKendall(GaringerOzone.monthly.ts)
summary(trend1)
```

```
## Score =  -77 , Var(Score) = 1499
## denominator =  539.4972
## tau = -0.143, 2-sided pvalue =0.046724
```

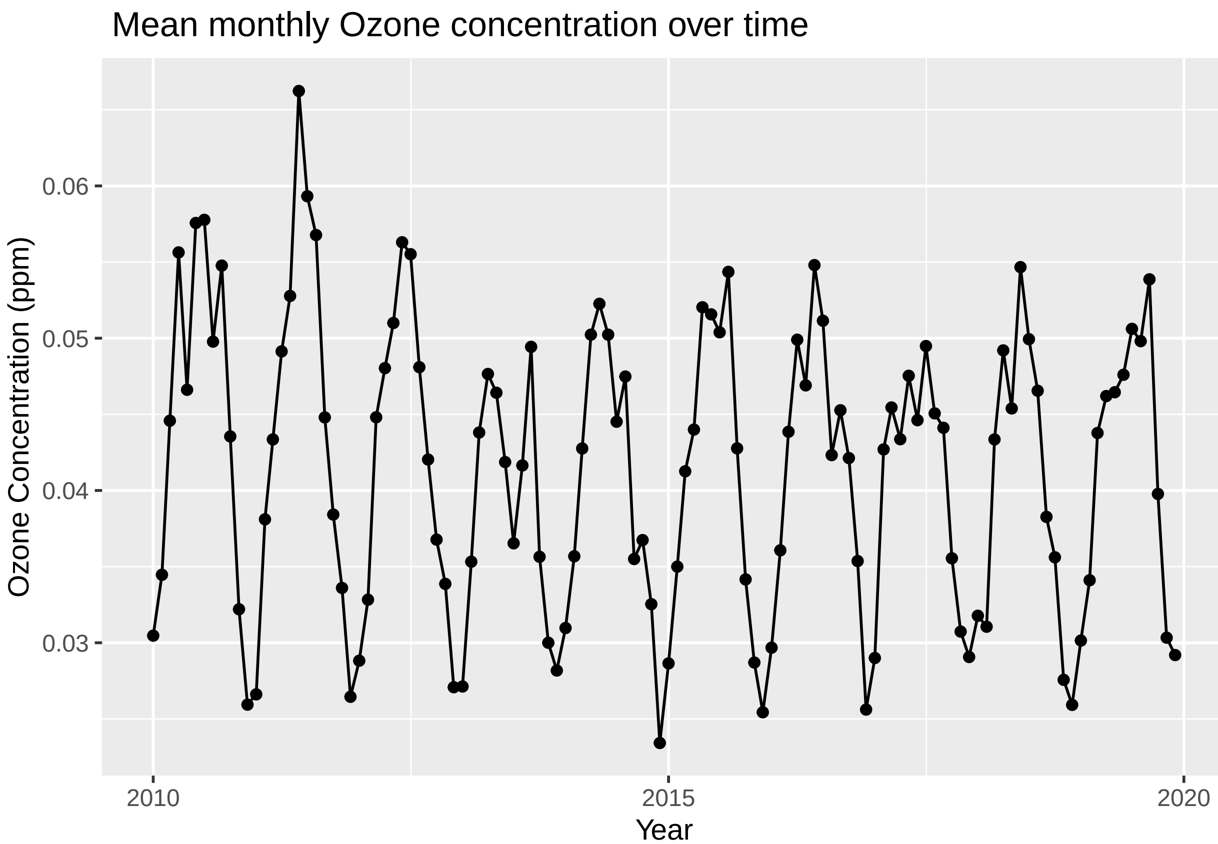
```r
trend2 <- trend::smk.test(GaringerOzone.monthly.ts)
summary(trend2)
```

```
##
##  Seasonal Mann-Kendall trend test (Hirsch-Slack test)
##
## data: GaringerOzone.monthly.ts
## alternative hypothesis: two.sided
##
## Statistics for individual seasons
##
## H0
##                      S varS    tau      z Pr(>|z|)
## Season 1:   S = 0   15  125  0.333  1.252  0.21050
## Season 2:   S = 0   -1  125 -0.022  0.000  1.00000
## Season 3:   S = 0   -4  124 -0.090 -0.269  0.78762
## Season 4:   S = 0  -17  125 -0.378 -1.431  0.15241
## Season 5:   S = 0  -15  125 -0.333 -1.252  0.21050
## Season 6:   S = 0  -17  125 -0.378 -1.431  0.15241
## Season 7:   S = 0  -11  125 -0.244 -0.894  0.37109
## Season 8:   S = 0   -7  125 -0.156 -0.537  0.59151
## Season 9:   S = 0   -5  125 -0.111 -0.358  0.72051
## Season 10:   S = 0 -13  125 -0.289 -1.073  0.28313
## Season 11:   S = 0 -13  125 -0.289 -1.073  0.28313
## Season 12:   S = 0  11  125  0.244  0.894  0.37109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: A man kendall test requires that our data does not have to meet the normality condition which is true for our case. It is a non parametric and is okay with no seasonality. We can use Mann Kendall test because we expect our data to have seasonal trned. Also this test can handle missing data.

13. Create a plot depicting mean monthly ozone concentrations over time, with both a geom_point and a geom_line layer. Edit your axis labels accordingly.

```r
# 13
Plot3 <- ggplot(GaringerOzone.monthly, aes(x = Date, y = MeanConcentrations)) +
  geom_line() +
  geom_point() +
  labs(x = "Year", y = "Ozone Concentration (ppm)", title=" Mean monthly Ozone concentration over time")
print(Plot3)
```



14. To accompany your graph, summarize your results in context of the research question. Include output from the statistical test in parentheses at the end of your sentence. Feel free to use multiple sentences in your interpretation.

    Answer: Our research question is whether there is trend in ozone concentration over time. Since all the p values are greater than 0.05 we cant reject the null hypothsis which states that there is no monotonic trend. All 12 seasons dont have a monotonic trend of ozone concentration over 2010s

15. Subtract the seasonal component from the `GaringerOzone.monthly.ts`. Hint: Look at how we extracted the series components for the EnoDischarge on the lesson Rmd file.

16. Run the Mann Kendall test on the non-seasonal Ozone monthly series. Compare the results with the ones obtained with the Seasonal Mann Kendall on the complete series.

```
# 15
ozone.seasonal.components <- as.data.frame(Decomposed_monthly$time.series[,1:3])%>%
  mutate(extracted = GaringerOzone.monthly$MeanConcentrations,
         Date = GaringerOzone.monthly$Date)

ozone.nonseasonal.components <- ts(data = ozone.seasonal.components$extracted - ozone.seasonal.componen
                                   start = c(2010,1),
                                   frequency = 12)

# 16
noneseasonal_trend <- Kendall::MannKendall(ozone.nonseasonal.components)
summary(noneseasonal_trend)
```

```
## Score =  -1179 , Var(Score) = 194365.7
```

```
## denominator =  7139.5
## tau = -0.165, 2-sided pvalue =0.0075402
```

Answer: The 2 sided p value is 0.0075402where we have removed the monthly seasonal component. This p value is significant and ndicates the monotonic trend. However, when we didnt remove the seasonality component in part 14 the p value was not signigicant. We can reject the null hypothesis in this case.