Assignment 7: Time Series Analysis

Natalie von Turkovich

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 Monday, March 14 at 7:00 pm.

Set up

1. Set up your session:

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

```
#1
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     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
```

```
#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
```

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

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

```
# Set theme
mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
theme_set(mytheme)
```

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.

```
#2


O3_GNC_19 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2019_raw.csv",
                      stringsAsFactors = TRUE)
O3_GNC_18 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2018_raw.csv",
                      stringsAsFactors = TRUE)
O3_GNC_17 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2017_raw.csv",
                      stringsAsFactors = TRUE)
O3_GNC_16 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2016_raw.csv",
                      stringsAsFactors = TRUE)
O3_GNC_15 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2015_raw.csv",
```

```r
                          stringsAsFactors = TRUE)
O3_GNC_14 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2014_raw.csv",
                          stringsAsFactors = TRUE)
O3_GNC_13 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2013_raw.csv",
                          stringsAsFactors = TRUE)
O3_GNC_12 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2012_raw.csv",
                          stringsAsFactors = TRUE)
O3_GNC_11 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2011_raw.csv",
                          stringsAsFactors = TRUE)
O3_GNC_10 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2010_raw.csv",
                          stringsAsFactors = TRUE)

  GaringerOzone <-rbind(O3_GNC_19, O3_GNC_18)
  GaringerOzone <-rbind(GaringerOzone, O3_GNC_17)
  GaringerOzone <-rbind(GaringerOzone, O3_GNC_16)
  GaringerOzone <-rbind(GaringerOzone, O3_GNC_15)
  GaringerOzone <-rbind(GaringerOzone, O3_GNC_14)
  GaringerOzone <-rbind(GaringerOzone, O3_GNC_13)
  GaringerOzone <-rbind(GaringerOzone, O3_GNC_12)
  GaringerOzone <-rbind(GaringerOzone, O3_GNC_11)
  GaringerOzone <-rbind(GaringerOzone, O3_GNC_10)
```

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.

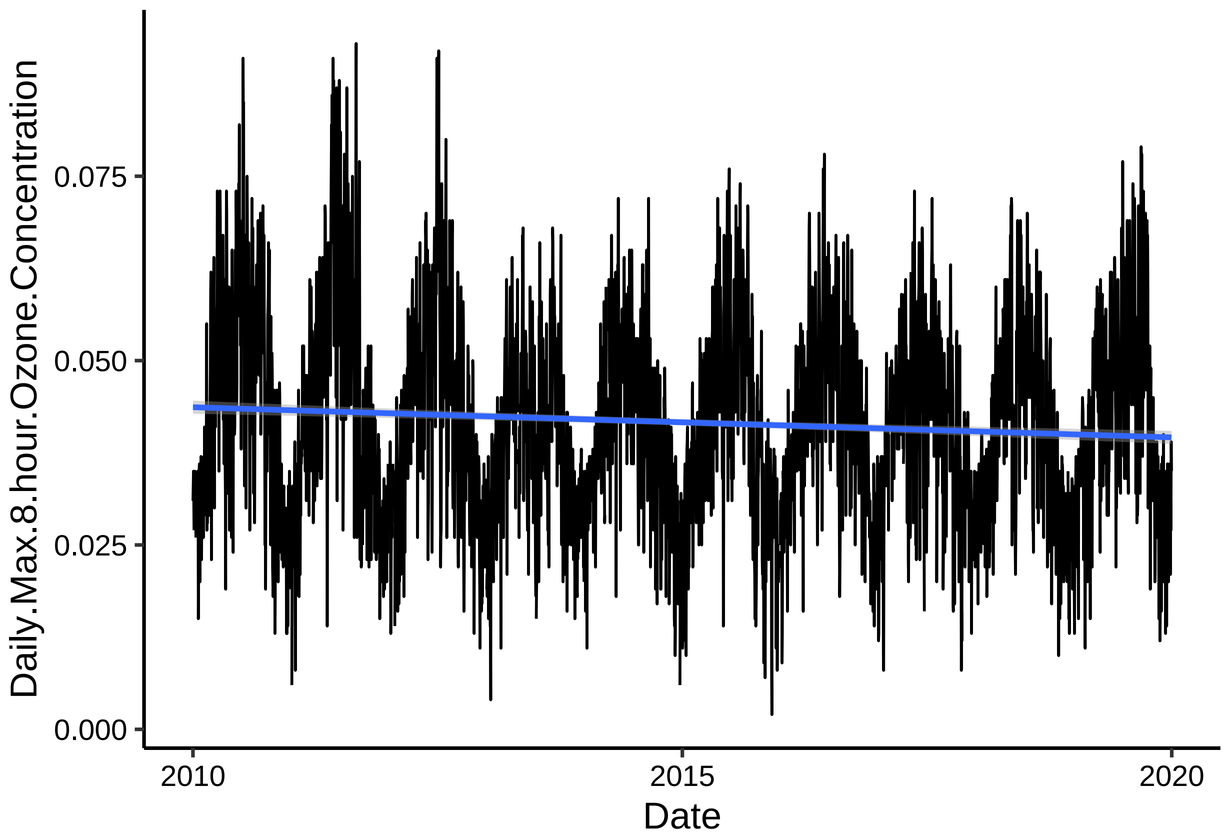
```r
# 3
GaringerOzone$Date <- mdy(GaringerOzone$Date)

# 4
GraingerOzone_clean <-GaringerOzone %>%
  select(c(Date, Daily.Max.8.hour.Ozone.Concentration, DAILY_AQI_VALUE))

# 5
Days_df <- data.frame(DATE = seq.Date(from = as.Date("2010-1-1"),
                                       to = as.Date("2019-12-31"), by = "day"))

Days_df  = rename(Days_df , Date = DATE)

# 6
GaringerOzone<-left_join(Days_df, GraingerOzone_clean, "Date")
```

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
GaringerOzone_plot <-
ggplot(GaringerOzone, aes(x = Date, y = Daily.Max.8.hour.Ozone.Concentration)) +
  geom_line() +
  geom_smooth( method = lm )
print(GaringerOzone_plot)
```

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

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



Answer: The plost suggests a decreasing trend in ozone concentration over time.

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? sp-quadratic, piecewise-equal to surrounds

```
#8
GaringerOzone_filled <-
 GaringerOzone %>%
 mutate(PPM.clean = zoo::na.approx(Daily.Max.8.hour.Ozone.Concentration) )
```

Answer: Spline interpolation uses a quadratic function to interpolate and piecewise fills missing data by assuming it to be equal to the measurement made nearest to that date which means it could be earlier or later. Linear interpolation is most sensible here as it will take the value inbetween the date beforeand the date after the missing date.

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_filled %>%
  mutate(month = lubridate::month(Date),
         year = lubridate::year(Date))%>%
  mutate(Date = my(paste0(month,"-",year))) %>%
  dplyr::group_by(Date, year, month) %>%
  dplyr::summarise(MeanOzone = mean(PPM.clean))
```

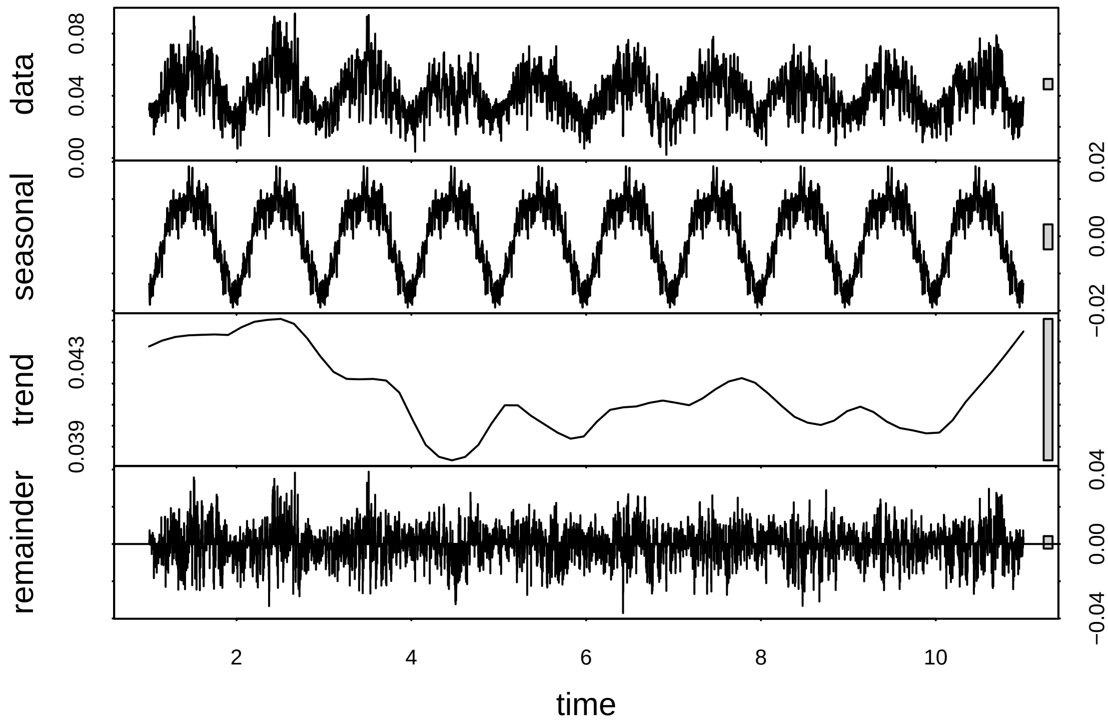
```
## `summarise()` has grouped output by 'Date', 'year'. You can override using the `.groups` argument.
```

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.

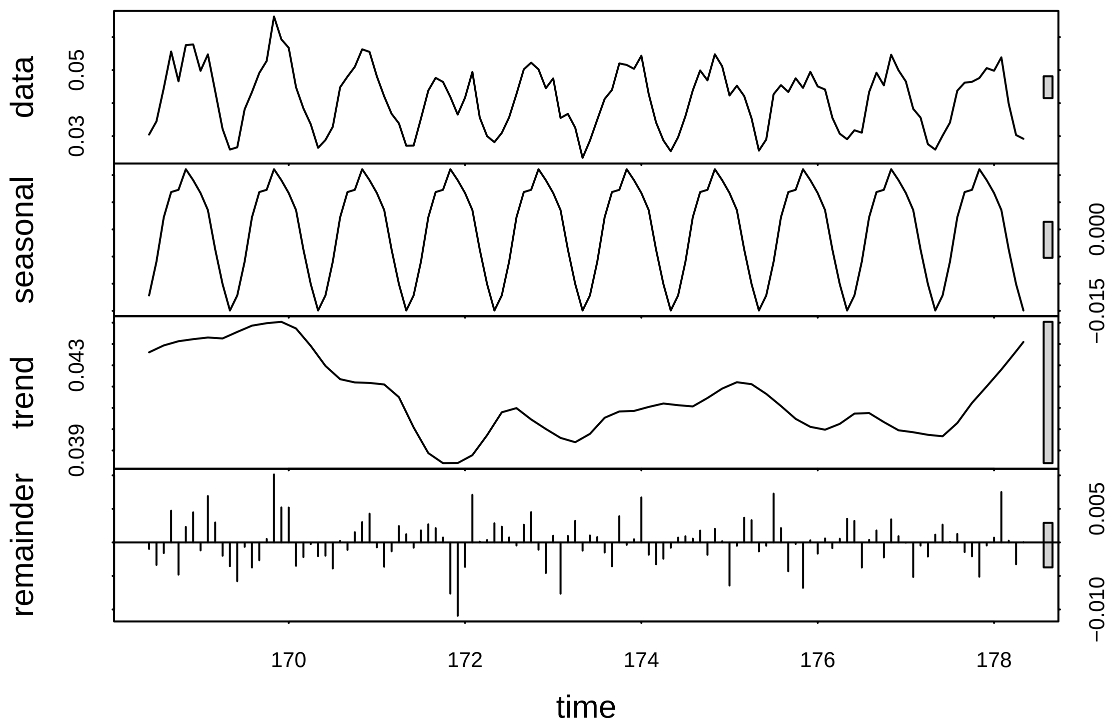
```
#10

fday_daily <- day(first(GaringerOzone_filled$Date))
fmonth_daily <- month(first(GaringerOzone_filled$Date))
fyear_daily <- year(first(GaringerOzone_filled$Date))

GaringerOzone.daily.ts <- ts(GaringerOzone_filled$PPM.clean,
        start =c(fday_daily,fmonth_daily, fyear_daily), frequency = 365)

fday_monthly <- day(first(GaringerOzone.monthly$Date))
fmonth_monthly <- month(first(GaringerOzone.monthly$Date))
fyear_monthly <- year(first(GaringerOzone.monthly$Date))

GaringerOzone.monthly.ts <- ts(GaringerOzone.monthly$MeanOzone,
          start = c(fmonth_monthly,fyear_monthly), frequency = 12)
```

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

```r
#11
GaringerOzone.daily.ts_Decomposed <-
  stl(GaringerOzone.daily.ts, s.window = "periodic") #pulls apart data
plot(GaringerOzone.daily.ts_Decomposed)
```




```r
GaringerOzone.monthly.ts_Decomposed <-
  stl(GaringerOzone.monthly.ts, s.window = "periodic") #pulls apart data
plot(GaringerOzone.monthly.ts_Decomposed)
```



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

Monotonic trends are a gradual shift over time that is consistent in direction. Specific tests for monotonic trend analysis are listed below, with assumptions and tips:

H0: S=0, ie. no trend H1: S!=0, ie., follow a trend

- **Seasonal Mann-Kendall**: seasonality, non-parametric `Kendall::SeasonalMannKendall` or `trend::smk.test()`

```
#12
O3_trend_monthly <- trend::smk.test(GaringerOzone.monthly.ts)
summary(O3_trend_monthly)
```

```
##
##  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   -7  125 -0.156 -0.537  0.59151
## Season 2:   S = 0   -5  125 -0.111 -0.358  0.72051
## Season 3:   S = 0  -13  125 -0.289 -1.073  0.28313
## Season 4:   S = 0  -13  125 -0.289 -1.073  0.28313
## Season 5:   S = 0   11  125  0.244  0.894  0.37109
## Season 6:   S = 0   15  125  0.333  1.252  0.21050
## Season 7:   S = 0   -1  125 -0.022  0.000  1.00000
## Season 8:   S = 0   -4  124 -0.090 -0.269  0.78762
## Season 9:   S = 0  -17  125 -0.378 -1.431  0.15241
## Season 10:   S = 0 -15  125 -0.333 -1.252  0.21050
## Season 11:   S = 0 -17  125 -0.378 -1.431  0.15241
## Season 12:   S = 0 -11  125 -0.244 -0.894  0.37109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
O3_trend_monthly.2 <-Kendall::SeasonalMannKendall(GaringerOzone.monthly.ts)
summary(O3_trend_monthly.2)
```

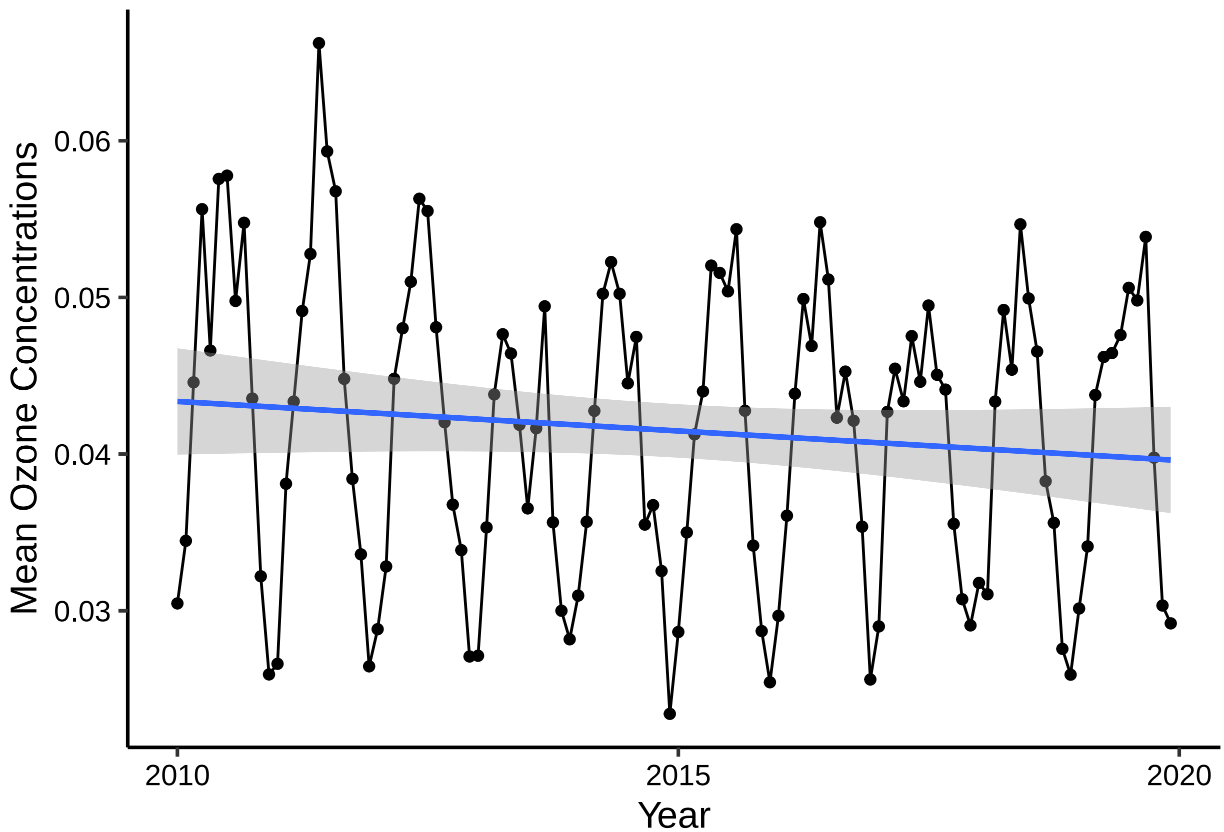
```
## Score =  -77 , Var(Score) = 1499
## denominator =  539.4972
## tau = -0.143, 2-sided pvalue =0.046724
```

> Answer: Mann Kendall seasonal, calculates a score for each month separately, gives a montly score so you can see whats happending month to month without being obstrcuted by whats happening in the months.

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.

```
# 13
GaringerOzone.monthly_plot <-
ggplot(GaringerOzone.monthly, aes(x = Date, y = MeanOzone)) +
  geom_point()+
  geom_line() +
  geom_smooth( method = lm )+
  labs(y="Mean Ozone Concentrations", x="Year")
print(GaringerOzone.monthly_plot)
```

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



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: The results of the seasonal Mann Kendall test result in a score of -77. We can reject the null hypothesis that S=0, there is evidence to support the alternate hypothsis of S!=0 (P<.05). The plot of the mean monthly PPM supports this test result as it depicts a decreasing trend in PPM over time.

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
GaringerOzone.monthly_components <-
as.data.frame(GaringerOzone.monthly.ts_Decomposed$time.series[,1:3])

GaringerOzone.monthly_components <-GaringerOzone.monthly_components %>%
  mutate(NonSeasonalOzone=
           GaringerOzone.monthly$MeanOzone-GaringerOzone.monthly_components$seasonal)

#16
O3_trend_NonSeasonal.2 <-Kendall::MannKendall(
```

```
  GaringerOzone.monthly_components$NonSeasonalOzone)
summary(O3_trend_NonSeasonal.2)
```

```
## Score =  -1179 , Var(Score) = 194365.7
## denominator =  7139.5
## tau = -0.165, 2-sided pvalue =0.0075402
```

> Answer: An analysis of the non-seasonal Ozone monthly series reveals a score of greater magnitude and of P=.0075. This P value is signficant (P<.05) and have a greater significance than the seasonal Mann Kenall test, indicating at subtracting out the seasonal component allows us to see the data trends more clearly.