Assignment 8: Time Series Analysis

Camber Vincent

Fall 2024

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on generalized linear models.

Directions

1. Rename this file `<FirstLast>_A08_TimeSeries.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change "Student Name" on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.

Set up

1. Set up your session:

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

```
getwd() #check working directory
```

```
## [1] "/Users/cambervincent/EDA_Fall_2024/Assignments"
```

```
#load necessary packages
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(zoo)
```

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

```
library(trend)
library(Kendall)

#set ggplot theme
theme_standard<-theme(

  text=element_text(family="Times",size=12,color="black"), #setting base font to Times New Roman
  plot.title=element_text(family="Helvetica",face="bold",size=16,  #title text theme
                          margin=margin(b=3)), #added margin for visual clarity
  plot.subtitle=element_text(family="Helvetica",face="italic",size=12, #subtitle text theme
                             color="gray20",
                             margin=margin(b=10)), #added margin for visual clarity

  plot.background=element_rect(fill="white"), #background set to white
  panel.background=element_rect(fill="white"), #graph background set to white
  panel.border=element_rect(color="black",fill=NA), #set a border around the graph

  panel.grid.major=element_line(color="gray85"), #recolor gridlines
  panel.grid.minor=element_line(color="gray95"),
  axis.ticks=element_blank() #turn off ticks

)

theme_set(theme_standard) #set custom theme as the base theme for all graphs
```

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.

```
#1
#import the ten datasets
#".." added to indicate up one directory from Assignments to main directory
ozone19<-read_csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2019_raw.csv")
ozone18<-read_csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2018_raw.csv")
ozone17<-read_csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2017_raw.csv")
ozone16<-read_csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2016_raw.csv")
ozone15<-read_csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2015_raw.csv")
ozone14<-read_csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2014_raw.csv")
ozone13<-read_csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2013_raw.csv")
ozone12<-read_csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2012_raw.csv")
ozone11<-read_csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2011_raw.csv")
```

```
ozone10<-read_csv("../Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2010_raw.csv")

#combine them into a single dataframe
GaringerOzone<-rbind(ozone19,ozone18,ozone17,ozone16,ozone15,ozone14,ozone13,ozone12,ozone11,ozone10)
dim(GaringerOzone) #check that dimensions match 3589 x 20
```

```
## [1] 3589   20
```

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$Date<-mdy(GaringerOzone$Date) #convert date column to date class

#4
GaringerOzone<-GaringerOzone%>%
  select(Date,'Daily Max 8-hour Ozone Concentration',DAILY_AQI_VALUE) #wrangle dataset

#5
Days<-as.data.frame(seq(as.Date("2010-01-01"),as.Date("2019-12-31"),by="day")) #create data frame
colnames(Days)<-c("Date") #rename column to Date

#6
GaringerOzone<-left_join(Days,GaringerOzone,by="Date") #calling leftjoin
dim(GaringerOzone) #checking dimensions match 3652 x 3
```

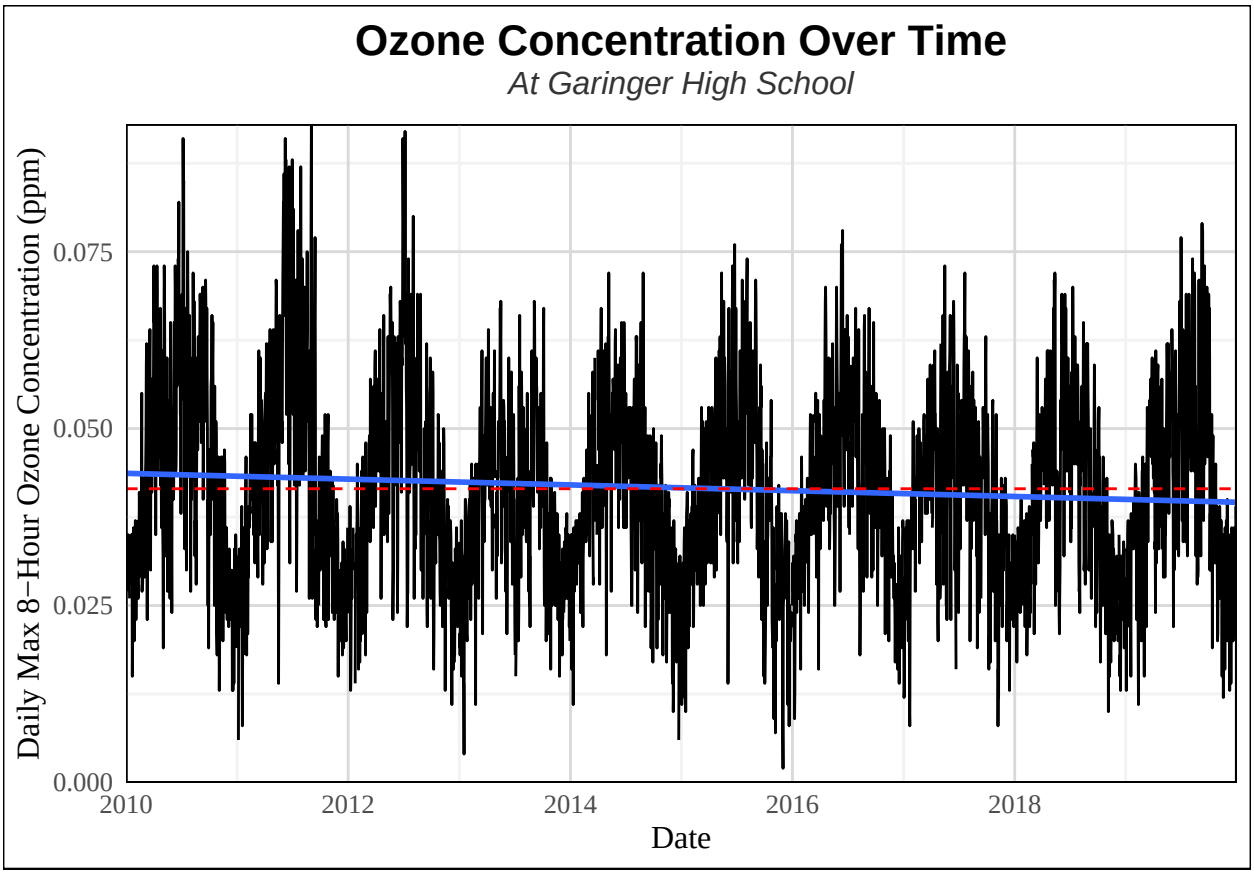
```
## [1] 3652    3
```

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
ggplot(GaringerOzone,aes(x=Date,y=`Daily Max 8-hour Ozone Concentration`))+ #create plot
  geom_line()+ #create line plot
  geom_smooth(method="lm",se=FALSE)+ #smoothed line w the confidence interval removed
```

```
#added horizontal line for visual comparison
geom_hline(yintercept=0.0415,linetype="dashed",color="red")+
labs(title="Ozone Concentration Over Time",
     subtitle="At Garinger High School",
     x="Date",
     y="Daily Max 8-Hour Ozone Concentration (ppm)")+ #labels

#setting axes limits and fixing 0,0 to bottom-left corner
scale_x_date(limits=c(as.Date("2010-01-01"),as.Date("2019-12-31")),expand=c(0,0))+
scale_y_continuous(limits=c(0,NA),expand=c(0,0))
```



Answer: The line plot does not suggest a strong trend in ozone concentration over time. The plotted smoothed line appears to be slightly decreasing over time, but is barely visually distinct from a straight line (plotted in red for reference above). Further non-visual analysis is needed to appropriately assess if there is a trend in ozone concnetration 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?

```
#8
GaringerOzoneClean<-GaringerOzone%>% #apply linear interpolation to fill in missing daily data
  mutate(ozone.clean=
           na.approx(`Daily Max 8-hour Ozone Concentration`,x=Date))
```

Answer: A linear interpolation is a straightfoward estimation that assumes a steady, gradual change between known data points which is appropriate for environmental data where gradual change is expected (like daily ozone levels). A piecewise constant interpolation would only carry forward/backward the last observed value which could misrepresent the ozone concentration trends since it doesn't account for gradual change. A piecewise interpolation would assume ozone levels stay consistent between observations, which is likely unrealistic. A spline interpretation creates complex curved shapes in interpolated data, which might produce an exaggerated trend in the ozone values which appear to vary back and forth over time quickly.

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)

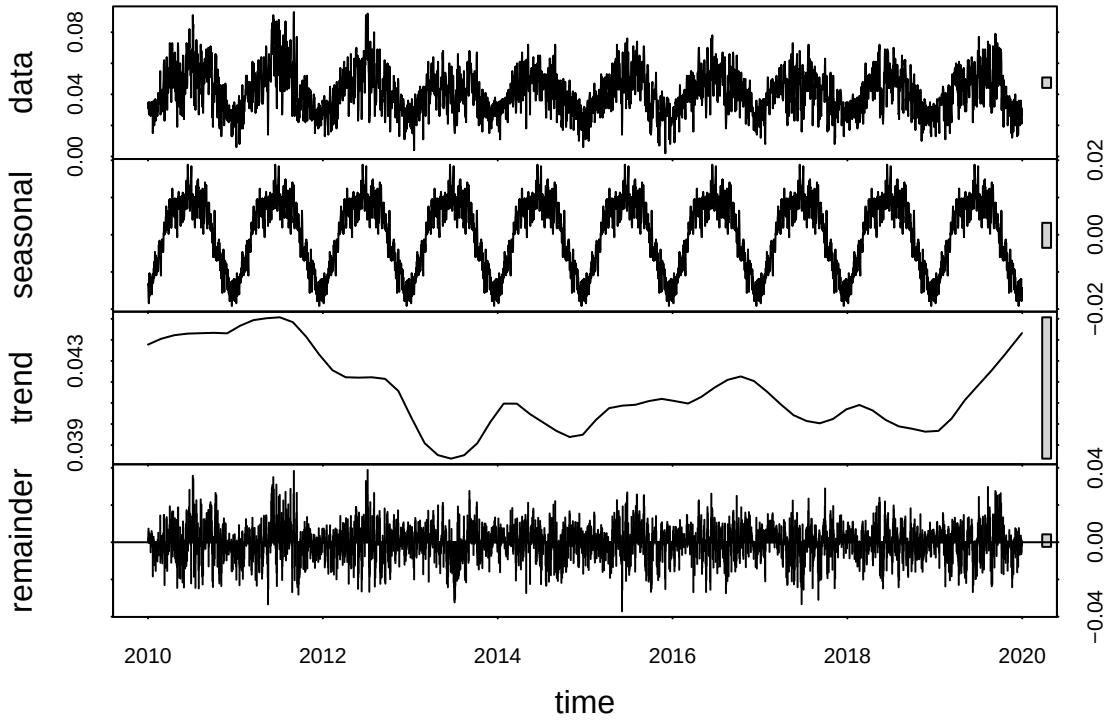
```
#9
GaringerOzone.monthly<-GaringerOzoneClean%>%
  mutate(Year=year(Date),Month=month(Date))%>% #added new columns for year and month
  group_by(Year,Month)%>% #group by year and month
  summarize(mean_ozone=mean(ozone.clean))%>% #perform mean data analysis
  ungroup()%>% #ungroup for further analysis
  #constructing date column in "2019-01-01" format
  mutate(Date=as.Date(paste(Year,Month,"01",sep="-")))
```

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.

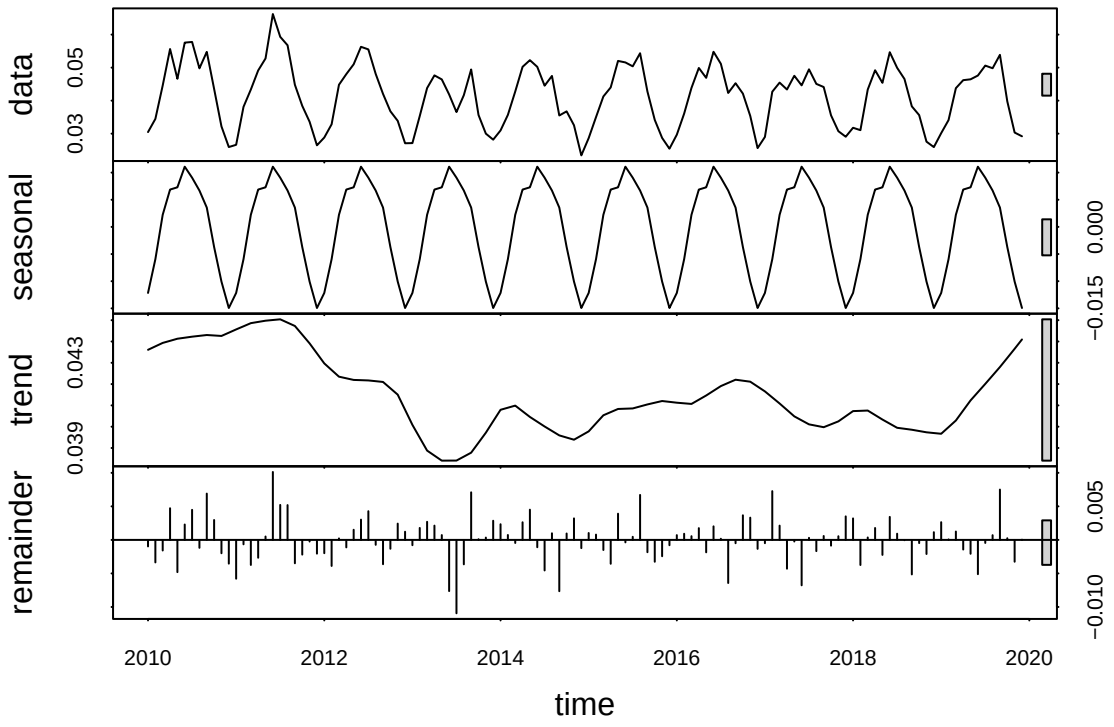
```
#10
GaringerOzone.daily.ts<-ts( #create daily observations time series object
  GaringerOzoneClean$ozone.clean,
  start=c(2010,1), #specify start date, year=2010 day=1
  end=c(2019,365), #specify end date
  frequency=365 #specify frequency
)

GaringerOzone.monthly.ts<-ts( #create monthly observations time series object
  GaringerOzone.monthly$mean_ozone,
  start=c(2010,1), #specify start date
  end=c(2019,12), #specify end date
  frequency=12 #specify frequency
)
```

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

```r
#11
GaringerOzone.dailydecomposed<-stl(GaringerOzone.daily.ts,s.window="periodic") #decomposition
plot(GaringerOzone.dailydecomposed) #plot the decomposition
```




```r
GaringerOzone.monthlydecomposed<-stl(GaringerOzone.monthly.ts,s.window="periodic") #decomposition
plot(GaringerOzone.monthlydecomposed) #plot the decomposition
```



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

```
#12
ozone_monthly_trend<-SeasonalMannKendall(GaringerOzone.monthly.ts) #run monotonic trend analysis
summary(ozone_monthly_trend) #inspect results
```

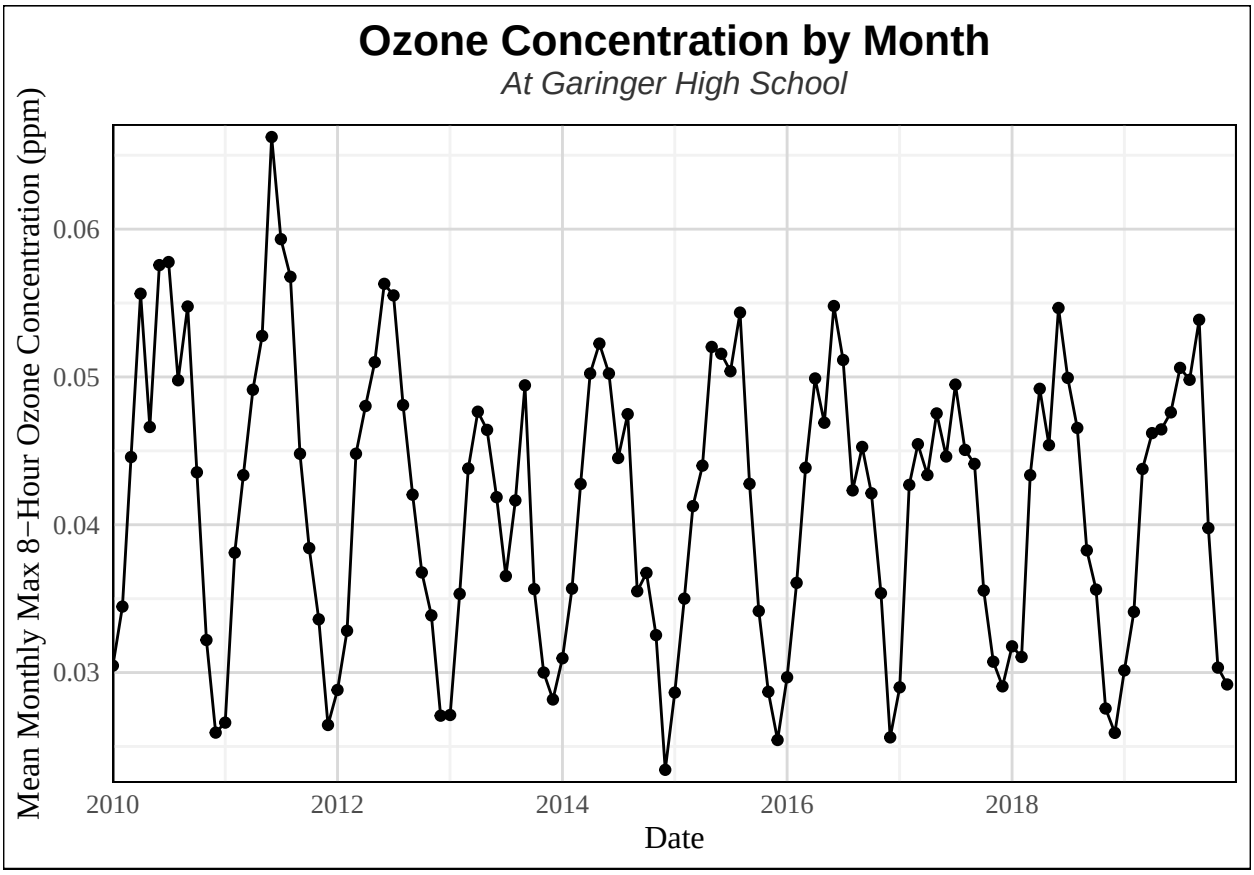
```
## Score =  -77 , Var(Score) = 1499
## denominator =  539.4972
## tau = -0.143, 2-sided pvalue =0.046724
```

Answer: The seasonal Mann-Kendall is most appropriate for the monotonic trend analysis as it is our only option (compared to linear regression, Mann-Kendall, Spearman Rho, and Augmented Dicky Fuller) that accurately acccounts for seasonality in the data. Looking at the plots of the decomposed daily and monthly time series, it is clear that our data has a seasonal variation that should be accounted for in the monotonic trend analysis.

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
ggplot(GaringerOzone.monthly,aes(x=Date,y=mean_ozone))+
  geom_point()+
  geom_line()+
```

```
labs(title="Ozone Concentration by Month",
     subtitle="At Garinger High School",
     x="Date",
     y="Mean Monthly Max 8-Hour Ozone Concentration (ppm)")+ #edit labels
scale_x_date(limits=c(as.Date("2010-01-01"),as.Date("2019-12-31")),expand=c(0,0))+
scale_y_continuous(expand=c(0.02,0)) #modify graph bounds for visual clarity
```



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 research question at hand is "have ozone concentrations changed over the 2010s at the Garinger High School station?" The results of the statistical test support the conclusion that yes, ozone concentrations have changed over time. The tau score of -0.143 indicates a slight decrease in ozone levels with a p-value of 0.047 falling below the significance threshold of 0.05, indicating a significant result that is unlikely to be due to random variation alone. (Score = -77, Var(Score) = 1499, denominator = 539.4972, tau = -0.143, 2-sided p-value = 0.046724)

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.

```r
#15
#create data frame from decomposed series
GaringerOzone.monthlycomponents<-as.data.frame(GaringerOzone.monthlydecomposed$time.series[,1:3])

#add original observed data and dates to the new table
GaringerOzone.monthlycomponents<-GaringerOzone.monthlycomponents%>%
  mutate(
    observed=GaringerOzone.monthly$mean_ozone, #adding observed values
    date=GaringerOzone.monthly$Date #adding date values
  )

#subtract the seasonal component from the original monthly series
GaringerOzone.monthlycomponents<-GaringerOzone.monthlycomponents%>%
  mutate(noseason=observed-seasonal)

#16
noseason_ozone<-GaringerOzone.monthlycomponents$noseason #extract ozone series with no seasons
ozone_noseason_trend<-MannKendall(noseason_ozone) #run monotonic trend analysis
summary(ozone_noseason_trend) #inspect results
```

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

Answer: The deseasonalized ozone Mann Kendall test indicates a stronger negative trend in ozone concentrations over time at Garinger High School than the Seasonal Mann Kendall test (tau = -0.165 < -0.143). Additionally, the results of the MK are more significant than the SMK test (p = 0.0075 < 0.047), falling beneath the 0.01 level of significance while the SMK test only fell beneath the 0.05 significance level. This result demonstrates that the overall decline in ozone levels is not a result of seasonal effects, and that the underlying trend in true decline is actually stronger when seasonal variations are taken into account.