Assignment 8: Time Series Analysis

Janice Ye

Spring 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

```
#1
getwd()
```

```
## [1] "/home/guest/EDA_Spring2024"
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## 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)

mytheme <- theme_classic(base_size = 12) +
  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
ozone10 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2010_raw.csv",
                     stringsAsFactors = T)
ozone11 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2011_raw.csv",
                     stringsAsFactors = T)
ozone12 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2012_raw.csv",
                     stringsAsFactors = T)
ozone13 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2013_raw.csv",
                     stringsAsFactors = T)
ozone14 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2014_raw.csv",
                     stringsAsFactors = T)
ozone15 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2015_raw.csv",
                     stringsAsFactors = T)
ozone16 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2016_raw.csv",
                     stringsAsFactors = T)
ozone17 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2017_raw.csv",
                     stringsAsFactors = T)
ozone18 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2018_raw.csv",
                     stringsAsFactors = T)
ozone19 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2019_raw.csv",
                     stringsAsFactors = T)

GaringerOzone <- rbind(ozone10, ozone11, ozone12, ozone13, ozone14, ozone15, ozone16, ozone17, ozone18,
```

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")
```

```r
# 4
GaringerOzone_short <- select(GaringerOzone, Date, Daily.Max.8.hour.Ozone.Concentration, DAILY_AQI_VALU

# 5
Days <- as.data.frame(seq(as.Date("2010-01-01"), as.Date("2019-12-31"), by = "days"))
names(Days) <- "Date"

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

```
## Joining with `by = join_by(Date)`
```

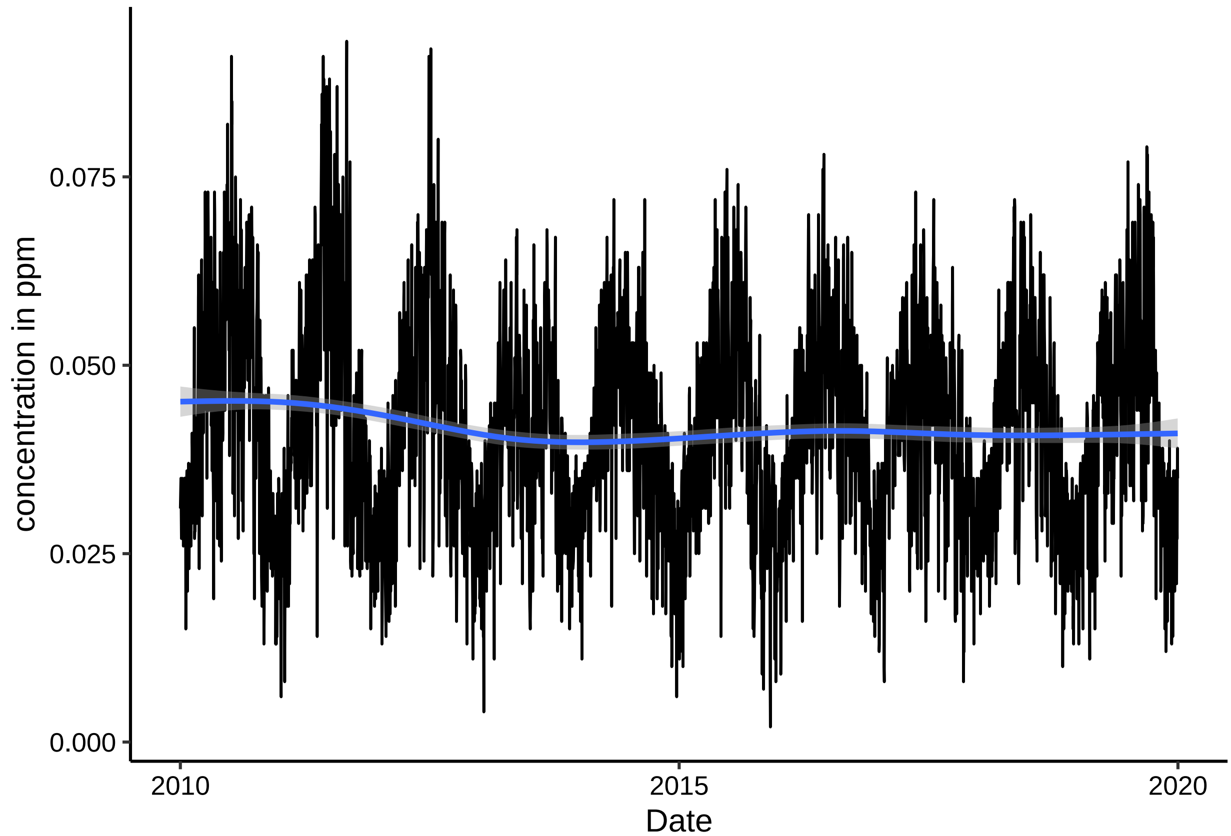
Visualize

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?

```r
#7
ozone_conc <- ggplot(GaringerOzone_short,
                     aes(x = Date, y = Daily.Max.8.hour.Ozone.Concentration)) +
  geom_line() +
  geom_smooth() +
  ylab("concentration in ppm")

print(ozone_conc)
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Answer: The trend is not very obvious but it is gradually decreasing over time.

Time Series Analysis

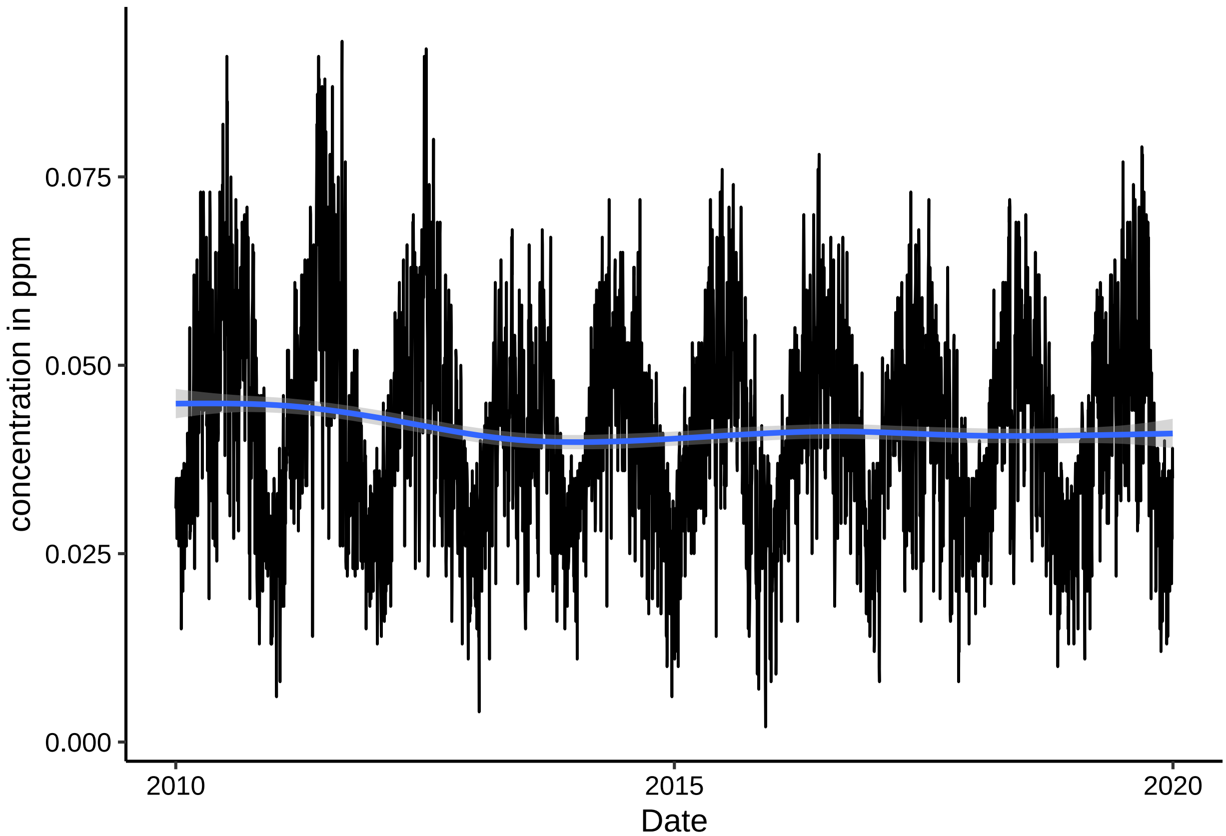
Study question: Have ozone concentrations changed over the 2010s at this station?

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
ozone.interpolation <- GaringerOzone %>%
  mutate(Daily.Max.8.hour.Ozone.Concentration = zoo::na.approx(Daily.Max.8.hour.Ozone.Concentration))

ggplot(ozone.interpolation, aes(x = Date, y = Daily.Max.8.hour.Ozone.Concentration)) +
  geom_line() +
  geom_smooth() +
  ylab("concentration in ppm")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Answer: The dataset does not have any short term of missing data, using spline interpolation might overshoot the data. Piecewise constant would be better for discontinuous data. Thus, we should use linear interpolation to approx data using a linear line.

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_short %>%
  mutate(Month = month(Date)) %>%
  mutate(Year = year(Date)) %>%
  mutate(Date = my(paste0(Month,"-",Year))) %>%
  group_by(Date) %>%
  mutate(Mean.Ozone.Concentration = mean(Daily.Max.8.hour.Ozone.Concentration)) %>%
  distinct(Date, Mean.Ozone.Concentration)
```

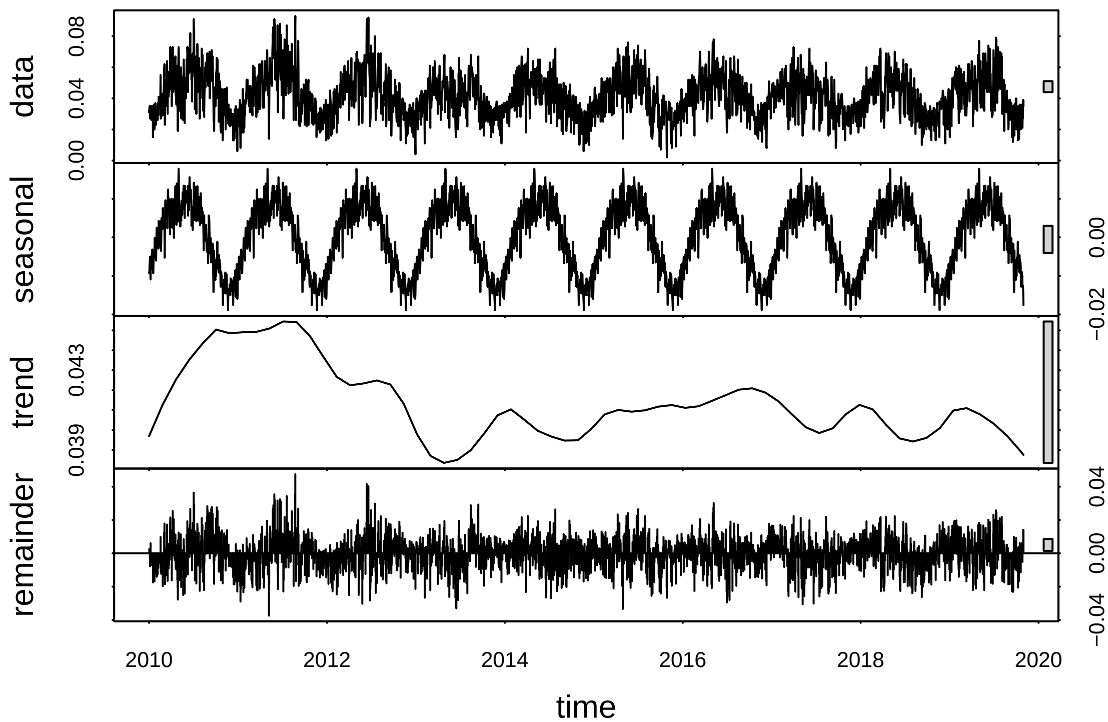
10. Generate two time series objects. Name the first `GaringerOzone.daily.ts` and base it on the dataframe of daily observations. Name the second `GaringerOzone.monthly.ts` and base it on the monthly average ozone values. Be sure that each specifies the correct start and end dates and the frequency of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

```
#10
GaringerOzone.daily.ts <- ts(GaringerOzone_short$Daily.Max.8.hour.Ozone.Concentration,
                   start = c(2010,1),
                   frequency = 365)
```

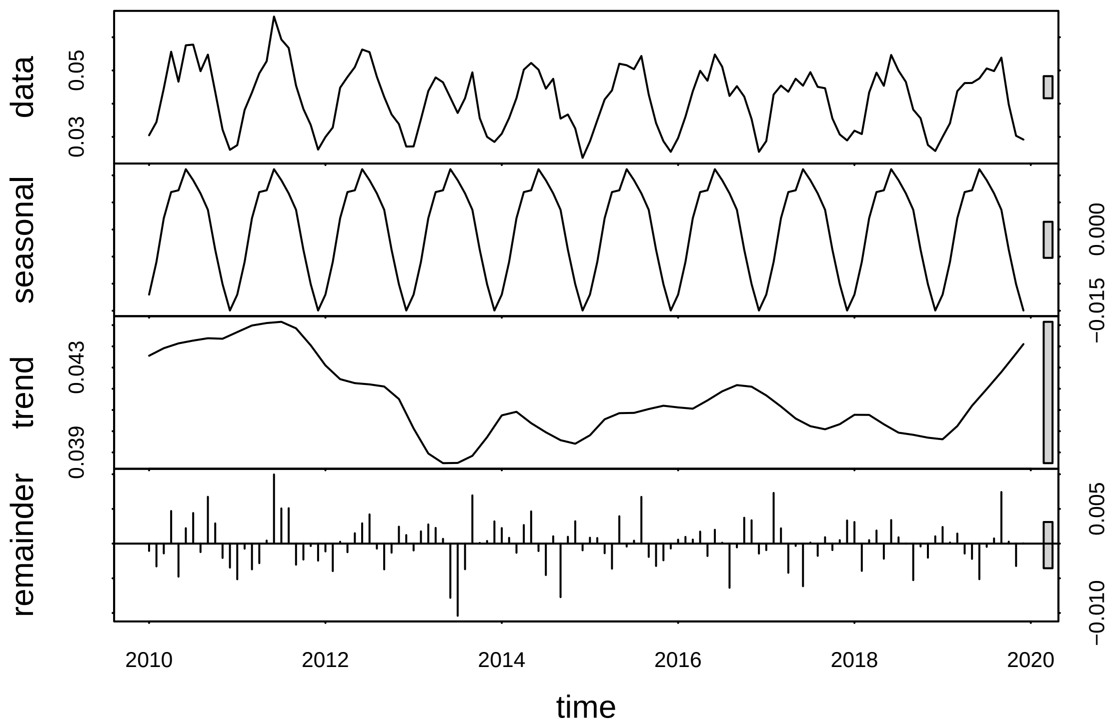
```r
GaringerOzone.monthly.ts <- ts(GaringerOzone.monthly$Mean.Ozone.Concentration,
                    start = c(2010,1),
                    frequency = 12)
```

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

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




```r
GaringerOzone.Monthly.Decomposed <- stl(GaringerOzone.monthly.ts, s.window = "periodic")
plot(GaringerOzone.Monthly.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?

```
#12
Monthly.Trend <- Kendall::SeasonalMannKendall(GaringerOzone.monthly.ts)
Monthly.Trend
```

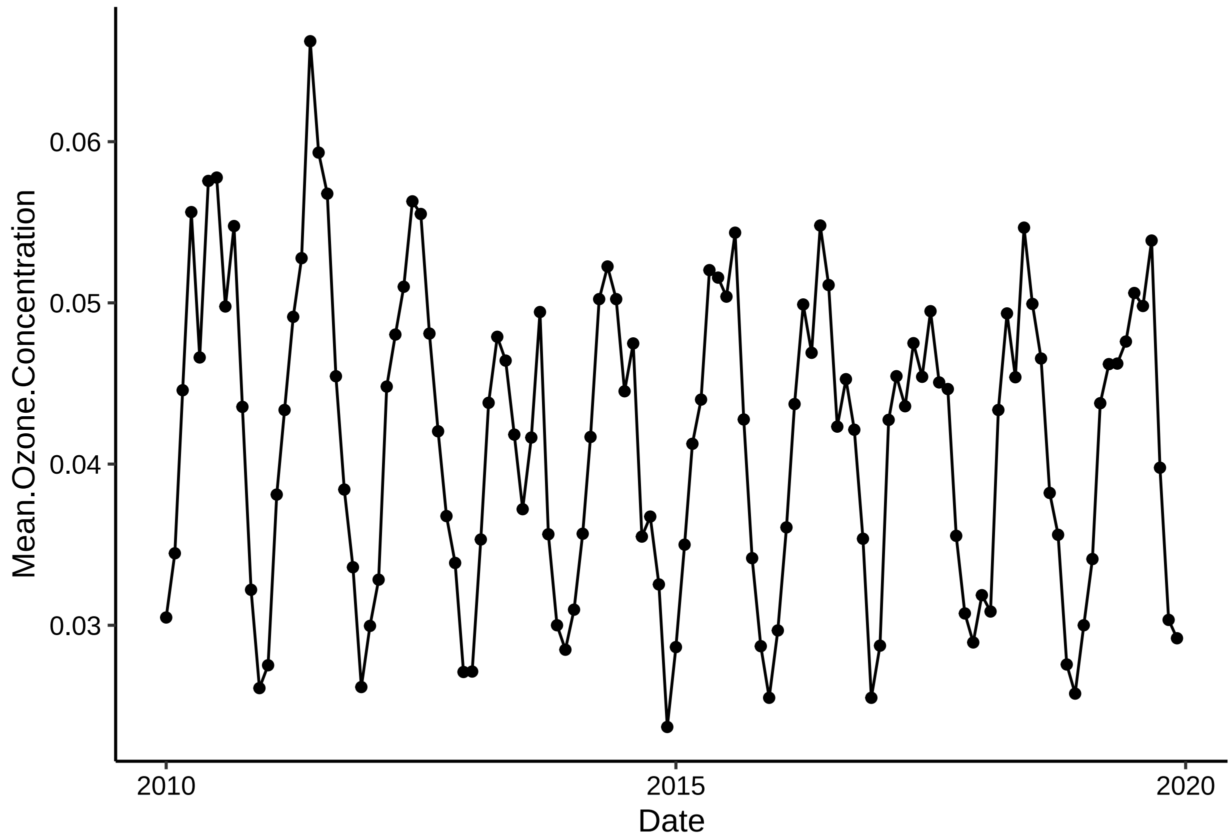
```
## tau = -0.163, 2-sided pvalue =0.022986
```

Answer: Seasonal Mann-Kendall is most appropriate because the data is monthly data, so smk can accounts for seasonal variation within the data. These variations might be overlooked if using other trend tools.

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
mean.monthly.ozone <- ggplot(GaringerOzone.monthly,
                             aes(x = Date, y = Mean.Ozone.Concentration)) +
  geom_point() +
  geom_line()

print(mean.monthly.ozone)
```



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 result is that ozone concentrations has changed over the 2010s at this station. Although the graph does not really show the trend, the test shows that ozone concentration has changed. The p-value is 0.02, which is smaller than 0.05. Since the null hypothesis of smk is that the data is stationary, we will reject the null hypothesis and indicate that there is a trend. (tau = -0.163, 2-sided pvalue =0.022986).

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.Components <- as.data.frame(GaringerOzone.Monthly.Decomposed$time.series[,1:3])

GaringerOzone.Components <- mutate(GaringerOzone.Components,
                                   Observed = GaringerOzone.monthly$Mean.Ozone.Concentration,
                                   Date = GaringerOzone.monthly$Date)

GaringerOzone.Nonseasonal <- GaringerOzone.monthly.ts - GaringerOzone.Components$seasonal
GaringerOzone.Nonseasonal
```

```
##             Jan        Feb        Mar        Apr        May        Jun
## 2010 0.04249527 0.04045030 0.04249632 0.04872154 0.03937511 0.04642715
```

```
## 2011 0.03953231 0.04409315 0.04127051 0.04222154 0.04553640 0.05509382
## 2012 0.04197379 0.03881360 0.04272212 0.04112154 0.04376221 0.04516049
## 2013 0.03914282 0.04130744 0.04171567 0.04098475 0.03918156 0.03068807
## 2014 0.04298153 0.04166458 0.03959749 0.04332154 0.04502027 0.03909382
## 2015 0.04065895 0.04098601 0.03917374 0.03708820 0.04479447 0.04042715
## 2016 0.04169121 0.04205744 0.04163981 0.04298820 0.03966544 0.04366049
## 2017 0.04074712 0.04872675 0.04336729 0.03667441 0.04026221 0.03427428
## 2018 0.04388046 0.03683786 0.04127051 0.04243303 0.03814931 0.04352715
## 2019 0.04201379 0.04009315 0.04168987 0.03928820 0.03899554 0.03646049
##             Jul        Aug        Sep        Oct        Nov        Dec
## 2010 0.04867697 0.04310746 0.05112456 0.04728154 0.04229487 0.04105167
## 2011 0.05022536 0.05010746 0.04180617 0.04215251 0.04369487 0.04111834
## 2012 0.04641890 0.04143004 0.03839123 0.04050735 0.04396154 0.04205167
## 2013 0.02810277 0.03497843 0.04579123 0.03937831 0.04009487 0.04343443
## 2014 0.03541890 0.04081714 0.03185789 0.04047509 0.04262820 0.03864133
## 2015 0.04128987 0.04768810 0.03912456 0.03789444 0.03879487 0.04045167
## 2016 0.04200992 0.03565585 0.04162456 0.04586219 0.04546154 0.04045167
## 2017 0.04038665 0.03839778 0.04101306 0.03928154 0.04082820 0.04388501
## 2018 0.04083826 0.03988165 0.03456623 0.03934606 0.03766154 0.04071029
## 2019 0.04151568 0.04313972 0.05022456 0.04350735 0.04042820 0.04414522
```

```r
#16
Nonseasonal.trend <- Kendall::MannKendall(GaringerOzone.Nonseasonal)
Nonseasonal.trend
```

```
## tau = -0.179, 2-sided pvalue =0.0037728
```

> Answer: The de-seasonalized version of data has a p-value of 0.004 (tau = -0.179, 2-sided pvalue = 0.0037728). Since the p-value is smaller than 0.05, we will reject the null hypothesis. Thus, there is change in data despite seasonal change.