Assignment 7: Time Series Analysis

Nicole Eastman

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.

```
# 1 Set-up
getwd()
```

```
## [1] "/home/guest/R/EDA Fall 2022"
```

```
library(tidyverse)
library(lubridate)
library(zoo)
library(trend)
library(dplyr)

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

# 2 Load Data
Ozone1 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2010_raw.csv",
    stringsAsFactors = TRUE)
Ozone2 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2011_raw.csv",
```

```
    stringsAsFactors = TRUE)
Ozone3 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2012_raw.csv",
    stringsAsFactors = TRUE)
Ozone4 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2013_raw.csv",
    stringsAsFactors = TRUE)
Ozone5 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2014_raw.csv",
    stringsAsFactors = TRUE)
Ozone6 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2015_raw.csv",
    stringsAsFactors = TRUE)
Ozone7 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2016_raw.csv",
    stringsAsFactors = TRUE)
Ozone8 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2017_raw.csv",
    stringsAsFactors = TRUE)
Ozone9 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2018_raw.csv",
    stringsAsFactors = TRUE)
Ozone10 <- read.csv("Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2019_raw.csv",
    stringsAsFactors = TRUE)
GaringerOzone <- rbind(Ozone1, Ozone2, Ozone3, Ozone4, Ozone5,
    Ozone6, Ozone7, Ozone8, Ozone9, Ozone10)
```

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 = "%d/%m/%Y")

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

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

colnames(Days) <- c("Date")

# 6
GaringerOzone.Joined <- left_join(GaringerOzone.Processed, Days,
    by = NULL, copy = FALSE, keep = FALSE)
```

```
## Joining, by = "Date"
```

Visualize

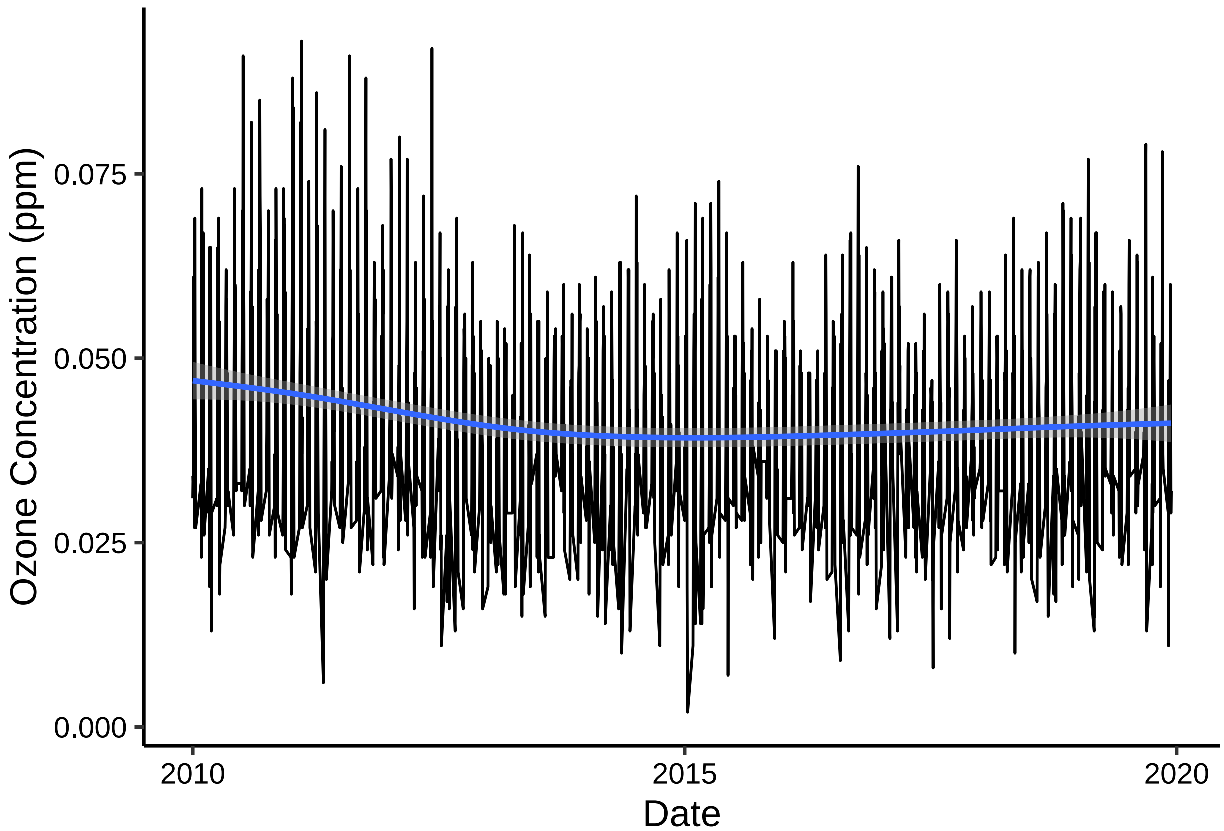
7. Create a line plot depicting ozone concentrations over time. In this case, we will plot actual concentrations in ppm, not AQI values. Format your axes accordingly. Add a smoothed line showing any linear trend of your data. Does your plot suggest a trend in ozone concentration over time?

```
# 7
Ozone.Plot <- ggplot(GaringerOzone.Joined, aes(x = Date, y = Daily.Max.8.hour.Ozone.Concentration)) +
    geom_line() + geom_smooth() + ylab("Ozone Concentration (ppm)")
print(Ozone.Plot)
```

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

## Warning: Removed 2168 rows containing non-finite values (stat_smooth).

## Warning: Removed 2168 row(s) containing missing values (geom_path).
```



Answer: There's not an obvious trend in ozone concentration over time. The smoothed line shows a decrease from 2010 to 2015 and then a slight increase from 2015 to 2020.

Time Series Analysis

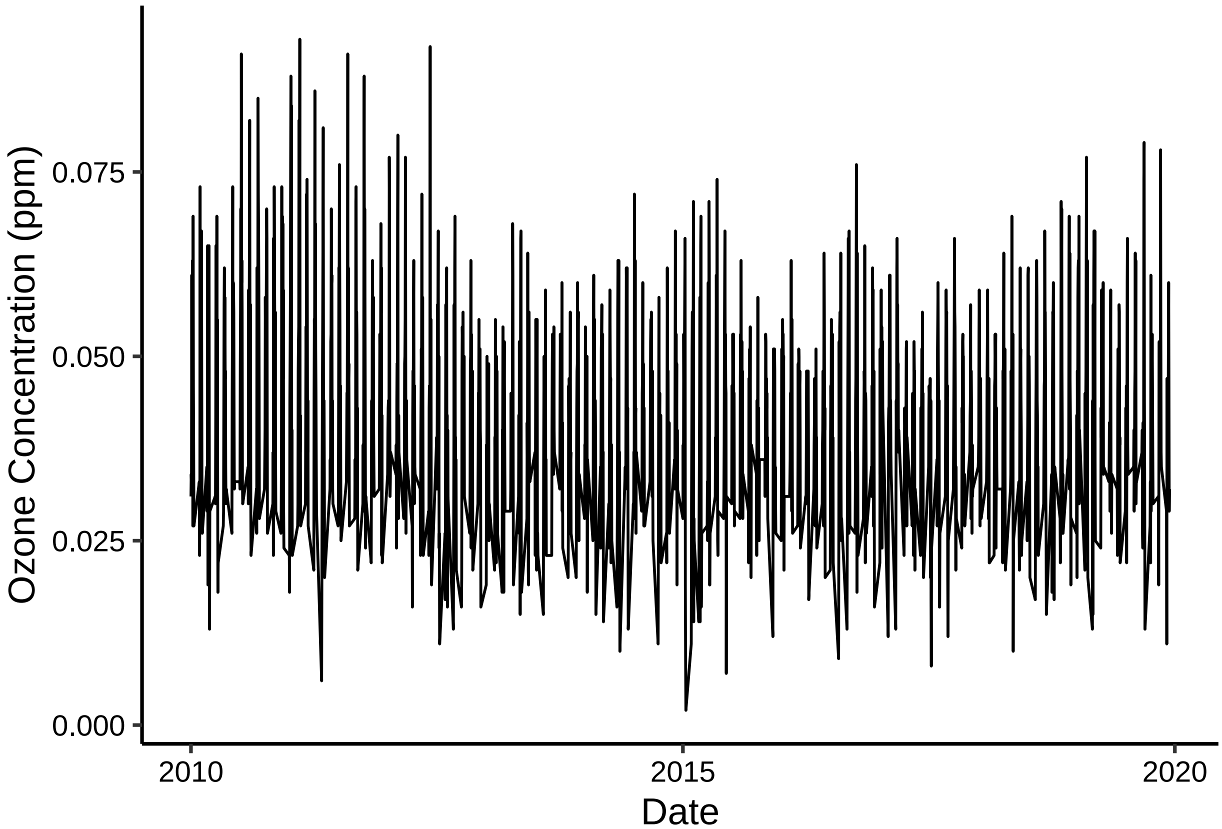
Study question: Have ozone concentrations changed over the 2010s at this station?

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

```
ggplot(Ozone.Clean) + geom_line(aes(x = Date, y = Daily.Max.8.hour.Ozone.Concentration)) +
    ylab("Ozone Concentration (ppm)")
```

```
## Warning: Removed 2168 row(s) containing missing values (geom_path).
```



Answer: We used the linear interpolation because the smoothed line in the previous questions shows mostly a linear trend. Additionally, the piecewise interpolation would match a value to the nearest data rather than continuing along the trend. The spline interpolation uses a quadratic function to fill-in the missing data, but there does not seem to be a exponential correlation between x and y for this dataframe.

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.Joined %>%
    mutate(Month = month(Date)) %>%
    mutate(Year = year(Date)) %>%
    mutate(month_year = my(paste0(Month, "-", Year))) %>%
    group_by(month_year) %>%
    summarize(mean_Ozone = mean(Daily.Max.8.hour.Ozone.Concentration))
```

```
## Warning: 2168 failed to parse.
```

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.

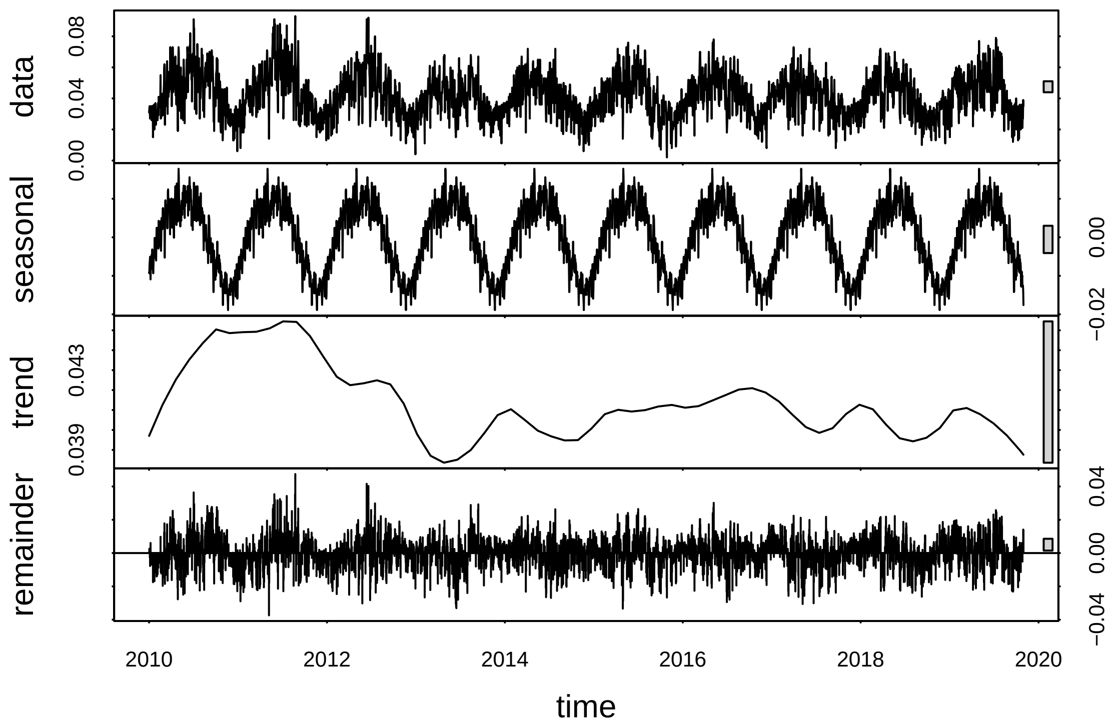
```
# 10
GaringerOzone.daily.ts <- ts(GaringerOzone.Joined$Daily.Max.8.hour.Ozone.Concentration,
    start = c(2010, 1), frequency = 365)

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

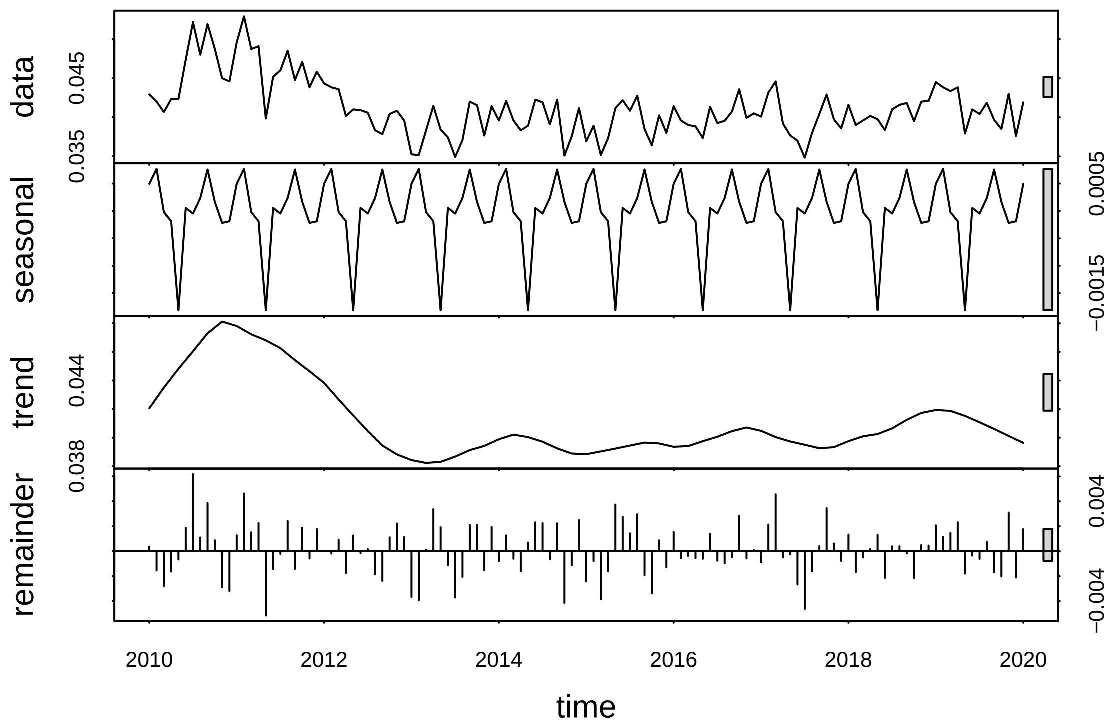
11. Decompose the daily and the monthly time series objects and plot the components using the `plot()` function.

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

plot(Ozone.Daily.Decomposed)
```




```
plot(Ozone.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
Ozone.Data.Trend <- Kendall::SeasonalMannKendall(GaringerOzone.monthly.ts)
Ozone.Data.Trend
```

```
## tau = -0.242, 2-sided pvalue =0.00069268
```

```
summary(Ozone.Data.Trend)
```

```
## Score =  -133 , Var(Score) = 1537
## denominator =  548.4916
## tau = -0.242, 2-sided pvalue =0.00069268
```

Answer: The Mann-Kendall is most appropriate because it doesn't require that the residuals be fitted to the parametric test like in linear regressions.

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.

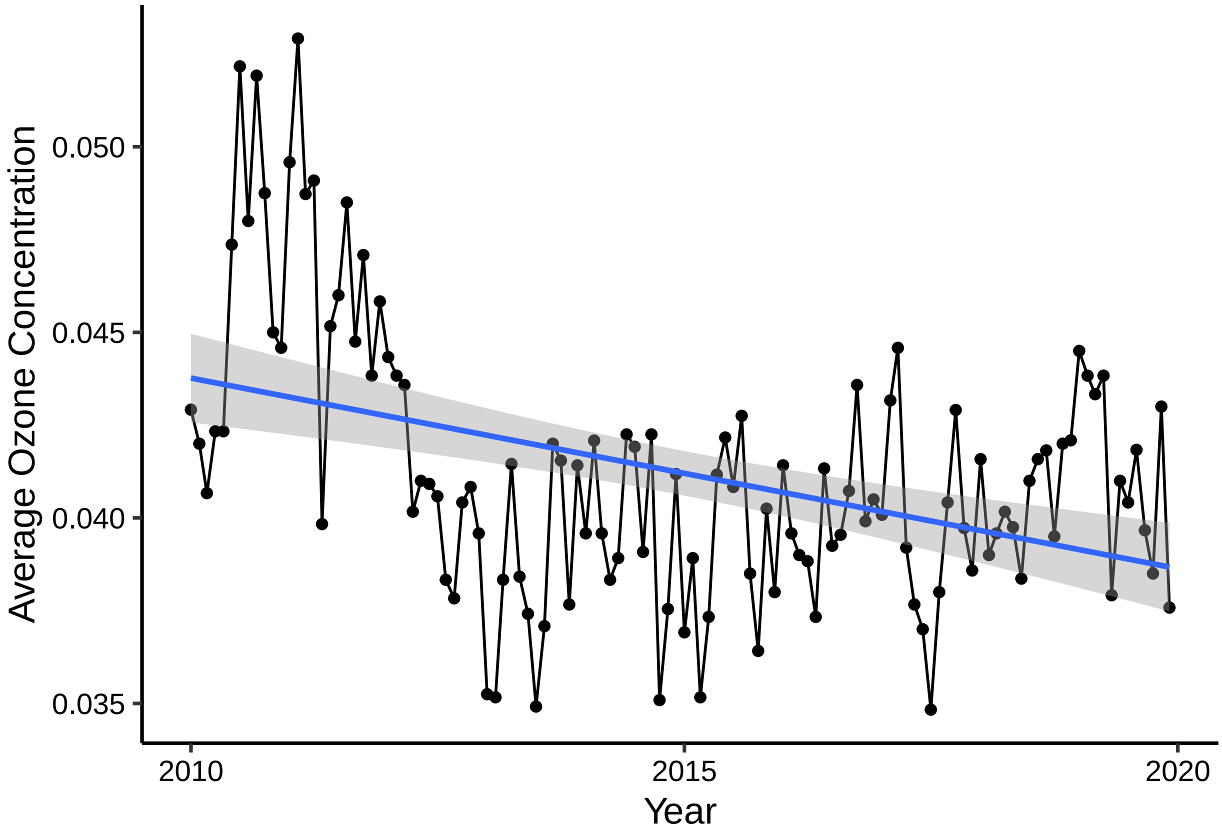
```
# 13
Ozone.MK.Plot <- ggplot(GaringerOzone.monthly, aes(x = month_year,
    y = mean_Ozone)) + geom_point() + geom_line() + xlab("Year") +
    ylab("Average Ozone Concentration") + geom_smooth(method = lm)
print(Ozone.MK.Plot)
```

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

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 1 row(s) containing missing values (geom_path).
```



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 average ozone concentration has decreased over time through the 2010s. There is a slight negative correlation between time and mean ozone concentation and the negative trend can be better explained by the correlation between the two variables (Date and Ozone Concentration) rather than chance. (tau= -0.242, 2-sided pvalue =0.0007)

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(Ozone.Monthly.Decomposed$time.series[,
    1:3])

Ozone.Seasonal.Components <- mutate(Ozone.Seasonal.Components,
    Observed = GaringerOzone.monthly$mean_Ozone, Date = GaringerOzone.monthly$month_year)

# 16
Ozone.Extracted.ts <- ts(Ozone.Seasonal.Components$Observed,
    start = c(-0.0018131, 1), frequency = 12)
```

```
Ozone.Extracted.Trend <- Kendall::SeasonalMannKendall(Ozone.Extracted.ts)
```

Answer: There is not much difference between the seasonal and non-seasonal monthly time series.