Assignment 7: Time Series Analysis

Yingchi Cheung

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
library(dplyr)
library(tidyverse)
library(lubridate)
library(trend)
library(zoo)
library(Kendall)
library(tseries)

mytheme <- theme_classic(base_size = 12) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "right")
theme_set(mytheme)

#2
EPAairFiles = list.files(path = "./Data/Raw/Ozone_TimeSeries/", pattern="*.csv", full.names=TRUE)
EPAairFiles
```

```
##  [1] "./Data/Raw/Ozone_TimeSeries//EPAair_O3_GaringerNC2010_raw.csv"
```

```
##  [2] "./Data/Raw/Ozone_TimeSeries//EPAair_O3_GaringerNC2011_raw.csv"
##  [3] "./Data/Raw/Ozone_TimeSeries//EPAair_O3_GaringerNC2012_raw.csv"
##  [4] "./Data/Raw/Ozone_TimeSeries//EPAair_O3_GaringerNC2013_raw.csv"
##  [5] "./Data/Raw/Ozone_TimeSeries//EPAair_O3_GaringerNC2014_raw.csv"
##  [6] "./Data/Raw/Ozone_TimeSeries//EPAair_O3_GaringerNC2015_raw.csv"
##  [7] "./Data/Raw/Ozone_TimeSeries//EPAair_O3_GaringerNC2016_raw.csv"
##  [8] "./Data/Raw/Ozone_TimeSeries//EPAair_O3_GaringerNC2017_raw.csv"
##  [9] "./Data/Raw/Ozone_TimeSeries//EPAair_O3_GaringerNC2018_raw.csv"
## [10] "./Data/Raw/Ozone_TimeSeries//EPAair_O3_GaringerNC2019_raw.csv"
```

```
GaringerOzone <- EPAairFiles %>%
  plyr::ldply(read.csv)
class(GaringerOzone)
```

```
## [1] "data.frame"
```

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")
class(GaringerOzone$Date)
```

```
## [1] "Date"
```

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

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



# 6
GaringerOzone1 <- left_join(Days, GaringerOzone_select)
```

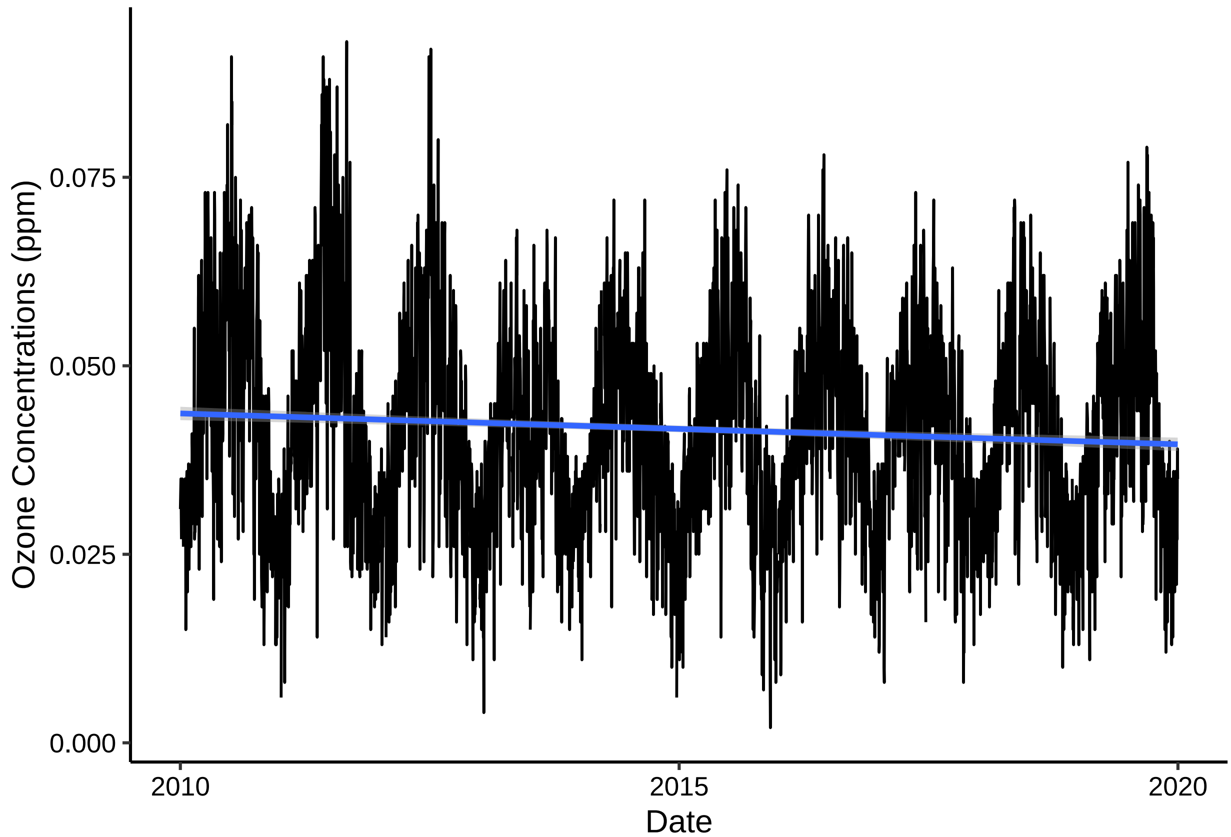
```
## Joining, by = "Date"
```

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?

```r
#7
GaringerOzone.line<-
  ggplot(GaringerOzone1, aes(x = Date, y = Daily.Max.8.hour.Ozone.Concentration)) +
  geom_line()+
  geom_smooth(method = "lm") +
  ylab("Ozone Concentrations (ppm)")
print(GaringerOzone.line)
```

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

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



Answer: My plot suggests there's 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?

```r
#8
summary(GaringerOzone1$Daily.Max.8.hour.Ozone.Concentration)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
## 0.00200 0.03200 0.04100 0.04163 0.05100 0.09300       63
```

```r
GaringerOzone1$Daily.Max.8.hour.Ozone.Concentration <- na.approx(GaringerOzone1$Daily.Max.8.hour.Ozone.C
```

```
summary(GaringerOzone1$Daily.Max.8.hour.Ozone.Concentration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00200 0.03200 0.04100 0.04151 0.05100 0.09300
```

> Answer: Piecewise constant will misrepresent the trend because missing data are assumed to be equal to the measurement made nearest to that date. Spline could make the trend difficult to see because it uses a quadratic function. Linear interpolation assumes missing data fall between the previous and next measurements, so the data won't be underestimated or overestimated. And a linear graph makes the trend easy to observe.

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 <-
  GaringerOzone1 %>%
  separate(Date, c("year", "month", "day")) %>%
  mutate(month = my(paste0(month, "-", year))) %>%
  group_by(month)%>%
  summarise (MeanOzone = mean(Daily.Max.8.hour.Ozone.Concentration))

colnames(GaringerOzone.monthly)[colnames(GaringerOzone.monthly) == "month"] = "Date"

print(GaringerOzone.monthly)
```

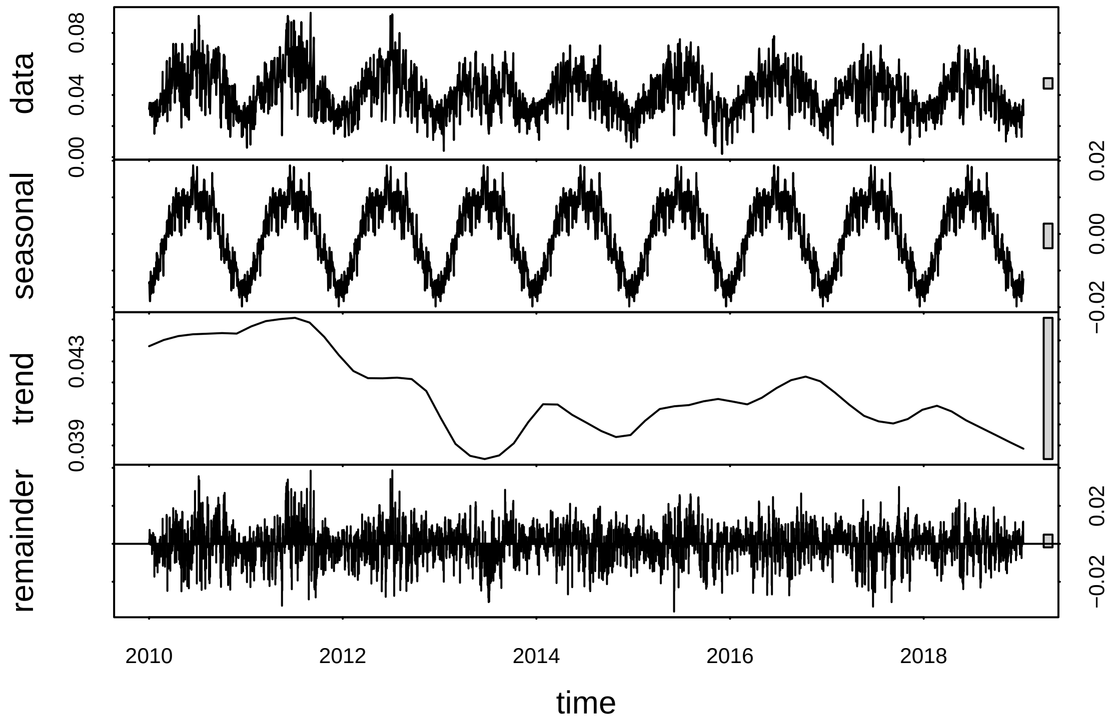
```
## # A tibble: 120 x 2
##    Date       MeanOzone
##    <date>         <dbl>
##  1 2010-01-01    0.0305
##  2 2010-02-01    0.0345
##  3 2010-03-01    0.0446
##  4 2010-04-01    0.0556
##  5 2010-05-01    0.0466
##  6 2010-06-01    0.0576
##  7 2010-07-01    0.0578
##  8 2010-08-01    0.0498
##  9 2010-09-01    0.0548
## 10 2010-10-01    0.0435
## # ... with 110 more rows
```

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.

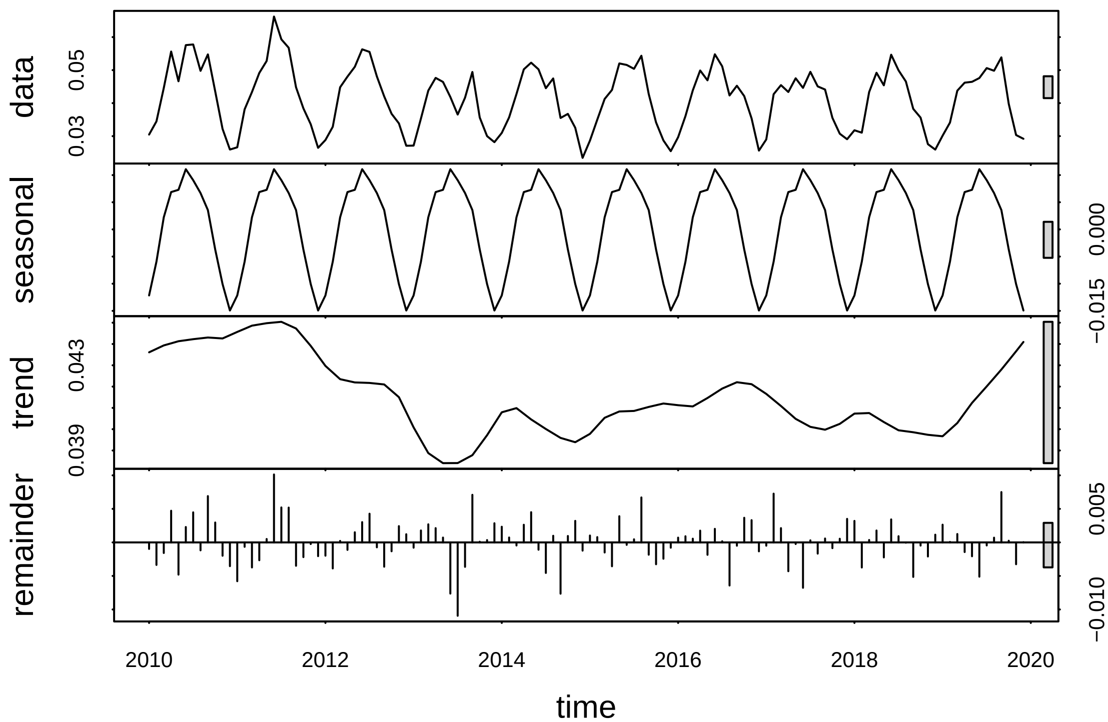
```
#10

GaringerOzone.daily.ts <- ts(GaringerOzone1$Daily.Max.8.hour.Ozone.Concentration, start=c(2010,01,01), 

GaringerOzone.monthly.ts <- ts(GaringerOzone.monthly$MeanOzone, start=c(2010,01,01), end=c(2019,12,31), 
```

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

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




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



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_trend1 <- Kendall::SeasonalMannKendall(GaringerOzone.monthly.ts)

Monthly_trend1
```

```
## tau = -0.143, 2-sided pvalue =0.046724
```

```
summary(Monthly_trend1)
```

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

```
Monthly_trend2 <- trend::smk.test(GaringerOzone.monthly.ts)

Monthly_trend2
```

```
##
##  Seasonal Mann-Kendall trend test (Hirsch-Slack test)
##
## data:  GaringerOzone.monthly.ts
## z = -1.963, p-value = 0.04965
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##    S varS
```

```
##  -77 1499
```

```
summary(Monthly_trend2)
```

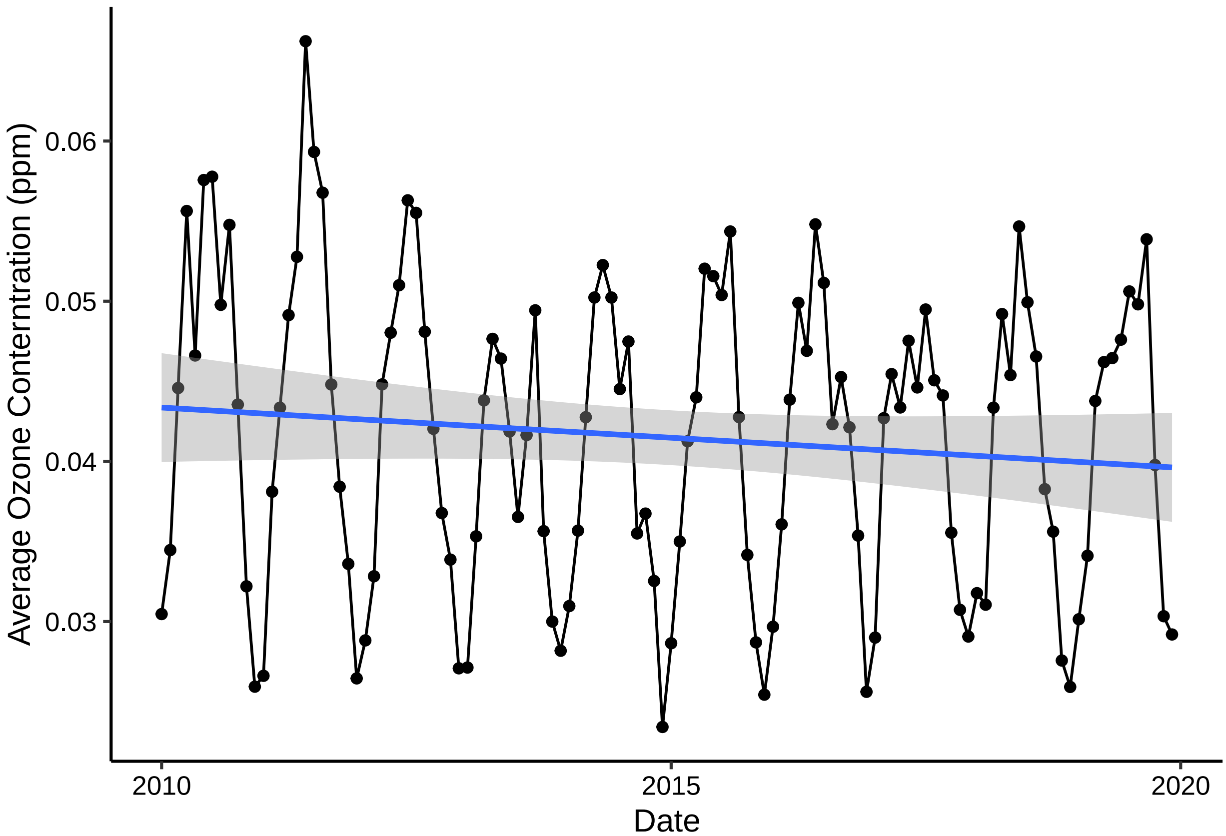
```
##
##  Seasonal Mann-Kendall trend test (Hirsch-Slack test)
##
## data: GaringerOzone.monthly.ts
## alternative hypothesis: two.sided
##
## Statistics for individual seasons
##
## H0
##                       S varS    tau      z Pr(>|z|)
## Season 1:   S = 0   15  125  0.333  1.252  0.21050
## Season 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: Seasonal cycles are present in the monthly ozone series. We want to see if ozone concentration has changed over the course of measurement while incorporating the seasonal component. To determine whether a monotonic trend exists in this situation, we use Seasonal Mann-Kendall test.

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
Monthly_trend_plot <-
ggplot(GaringerOzone.monthly, aes(x = Date, y = MeanOzone)) +
  geom_point() +
  geom_line() +
  ylab("Average Ozone Conterntration (ppm)") +
  geom_smooth(method = lm)
print(Monthly_trend_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: We can see that the trend line is decreasing from the mean monthly ozone concentrations over the time plot and the decomposed daily time series. Also, the S value from the Seasonal Mann-Kendall test shows an overall decreasing value for each season. Moreover, the p-value (0.049) is less than the alpha (0.05). we reject the null hypothesis, and the slope is not equal to 0. Therefore, the ozone concentrations changed over the 2010s at this station.

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
Monthly_components <- as.data.frame(Monthly_decomp$time.series[,1:3])

Nonseasonal <- mutate(Monthly_components,
        Observed = GaringerOzone.monthly$MeanOzone,
        Date = GaringerOzone.monthly$Date)



#16
Nonseasonal.ts <- ts(Nonseasonal$Observed, start = c(2010-01-01), frequency = 12)

Nonseasonal_trend <- trend::mk.test(Nonseasonal.ts)
```

```
Nonseasonal_trend
```

```
##
##  Mann-Kendall trend test
##
## data:  Nonseasonal.ts
## z = -0.95947, n = 120, p-value = 0.3373
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
## -4.240000e+02  1.943647e+05 -5.939207e-02
```

```
summary(Nonseasonal_trend)
```

```
##             Length Class  Mode
## data.name   1      -none- character
## p.value     1      -none- numeric
## statistic   1      -none- numeric
## null.value  1      -none- numeric
## parameter   1      -none- numeric
## estimates   3      -none- numeric
## alternative 1      -none- character
## method      1      -none- character
## pvalg       1      -none- numeric
```

> Answer: The p-value in the non-seasonal Ozone monthly series indicates that we fail to reject the null hypothesis because the p-value (0.3373) is greater than the alpha (0.05). The slope of the non-seasonal Ozone monthly series is equal to 0. For the non-seasonal Ozone monthly series, the ozone concentrations didn't change over the 2010s at this station. Compare to the Seasonal Mann Kendall on the complete series, they have the opposite outcome since the seasonal Ozone monthly series shows that ozone concentrations did change over the 2010s at this station.