Assignment 8: Time Series Analysis

Weilin Wang

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()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```
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
```

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

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

```r
library(trend)

theme_set(theme_minimal())
```

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.

```r
#1
path <- "~/EDE_Fall2024/Data/Raw/Ozone_TimeSeries/"

files <- list.files(path, pattern = "\\.csv$", full.names = TRUE)

GaringerOzone <- files %>%
  lapply(read.csv) %>%
  bind_rows() %>%
  mutate(Date = mdy(Date))

dim(GaringerOzone)
```

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

```r
head(GaringerOzone)
```

```
##         Date Source   Site.ID POC Daily.Max.8.hour.Ozone.Concentration UNITS
## 1 2010-01-01    AQS 371190041   1                                0.031   ppm
## 2 2010-01-02    AQS 371190041   1                                0.033   ppm
## 3 2010-01-03    AQS 371190041   1                                0.035   ppm
## 4 2010-01-04    AQS 371190041   1                                0.031   ppm
## 5 2010-01-05    AQS 371190041   1                                0.027   ppm
## 6 2010-01-07    AQS 371190041   1                                0.033   ppm
##   DAILY_AQI_VALUE            Site.Name DAILY_OBS_COUNT PERCENT_COMPLETE
## 1              29 Garinger High School              17              100
## 2              31 Garinger High School              17              100
## 3              32 Garinger High School              17              100
## 4              29 Garinger High School              17              100
## 5              25 Garinger High School              17              100
## 6              31 Garinger High School              17              100
##   AQS_PARAMETER_CODE AQS_PARAMETER_DESC CBSA_CODE
## 1              44201              Ozone     16740
```

```
## 2             44201             Ozone     16740
## 3             44201             Ozone     16740
## 4             44201             Ozone     16740
## 5             44201             Ozone     16740
## 6             44201             Ozone     16740
##                            CBSA_NAME STATE_CODE          STATE COUNTY_CODE
## 1 Charlotte-Concord-Gastonia, NC-SC         37 North Carolina         119
## 2 Charlotte-Concord-Gastonia, NC-SC         37 North Carolina         119
## 3 Charlotte-Concord-Gastonia, NC-SC         37 North Carolina         119
## 4 Charlotte-Concord-Gastonia, NC-SC         37 North Carolina         119
## 5 Charlotte-Concord-Gastonia, NC-SC         37 North Carolina         119
## 6 Charlotte-Concord-Gastonia, NC-SC         37 North Carolina         119
##        COUNTY SITE_LATITUDE SITE_LONGITUDE
## 1 Mecklenburg       35.2401      -80.78568
## 2 Mecklenburg       35.2401      -80.78568
## 3 Mecklenburg       35.2401      -80.78568
## 4 Mecklenburg       35.2401      -80.78568
## 5 Mecklenburg       35.2401      -80.78568
## 6 Mecklenburg       35.2401      -80.78568
```

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
head(GaringerOzone$Date)
```

```
## [1] "2010-01-01" "2010-01-02" "2010-01-03" "2010-01-04" "2010-01-05"
## [6] "2010-01-07"
```

```r
GaringerOzone$Date <- as.Date(GaringerOzone$Date)



# 4
GaringerOzone <- subset(GaringerOzone,
                        select = c(Date, Daily.Max.8.hour.Ozone.Concentration, DAILY_AQI_VALUE))

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

```r
# 6
GaringerOzone <- Days %>% left_join(GaringerOzone, by = "Date")
dim(GaringerOzone)
```

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

```r
head(GaringerOzone)
```

```
##         Date Daily.Max.8.hour.Ozone.Concentration DAILY_AQI_VALUE
## 1 2010-01-01                                0.031              29
## 2 2010-01-02                                0.033              31
## 3 2010-01-03                                0.035              32
## 4 2010-01-04                                0.031              29
## 5 2010-01-05                                0.027              25
## 6 2010-01-06                                   NA              NA
```

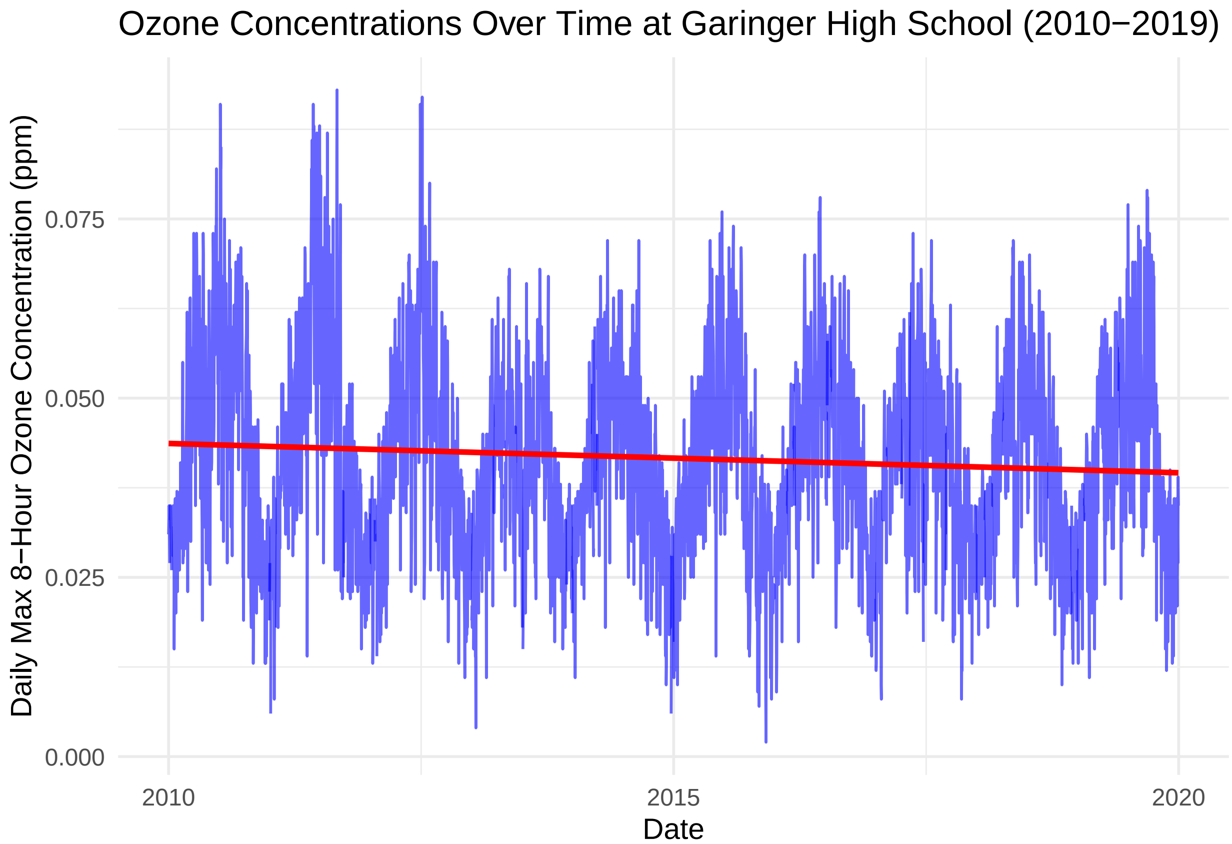
Visualize

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
ggplot(GaringerOzone, aes(x = Date, y = Daily.Max.8.hour.Ozone.Concentration)) +
  geom_line(color = "blue", alpha = 0.6) +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(
    title = "Ozone Concentrations Over Time at Garinger High School (2010-2019)",
    x = "Date",
    y = "Daily Max 8-Hour Ozone Concentration (ppm)"
  ) +
  theme_minimal()
```

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

```
## Warning: Removed 63 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



Answer:

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
GaringerOzone$Daily.Max.8.hour.Ozone.Concentration <- na.approx(GaringerOzone$Daily.Max.8.hour.Ozone.Con
```

Answer: Linear interpolation is used to fill missing ozone concentration data because it maintains overall trends without adding complexity or unrealistic fluctuations. Unlike piecewise constant interpolation, which creates abrupt changes, linear interpolation provides a gradual transition that better suits environmental data. Spline interpolation, though smoother, can introduce artificial oscillations and overfitting, especially with large gaps. Linear interpolation strikes a balance, preserving the trend integrity needed for analyzing changes over time.

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)

```r
#9
GaringerOzone.monthly <- GaringerOzone %>%
  mutate(year = year(Date), month = month(Date)) %>%
  group_by(year, month) %>%
  summarize(mean_ozone = mean(Daily.Max.8.hour.Ozone.Concentration, na.rm = TRUE)) %>%
  ungroup()
```

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

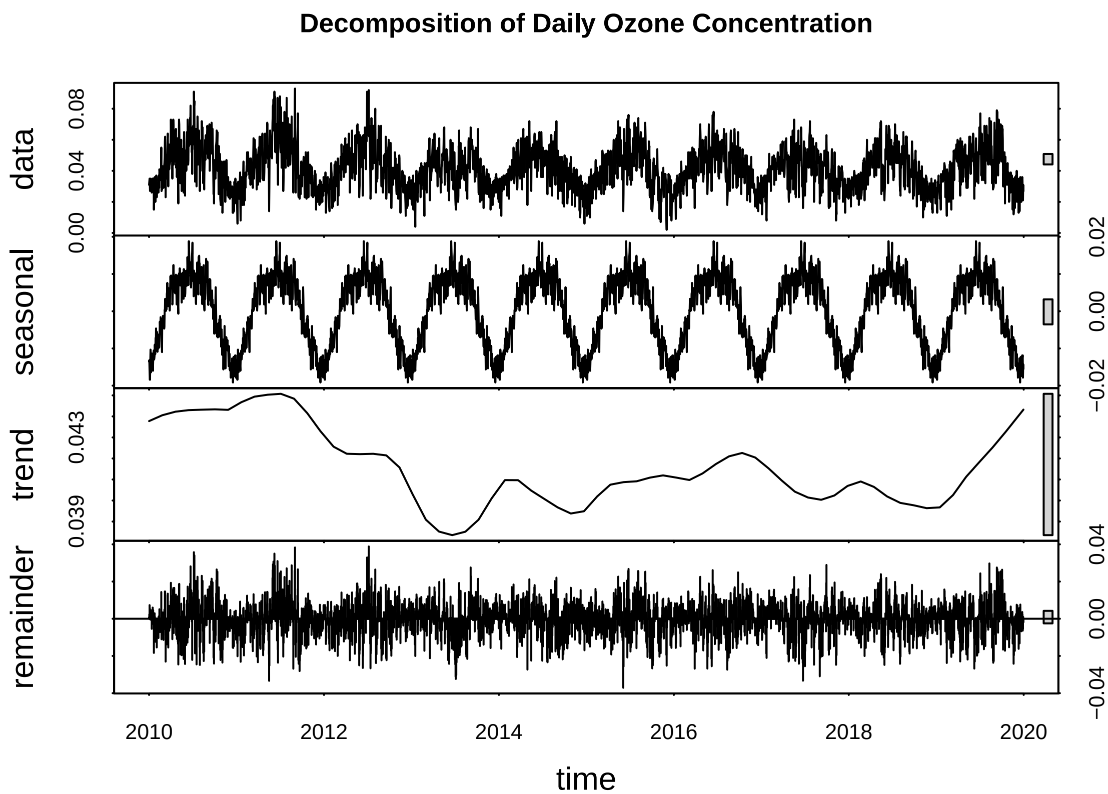
```r
GaringerOzone.monthly <- GaringerOzone.monthly %>%
  mutate(Date = as.Date(paste(year, month, "01", sep = "-")))
```

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.

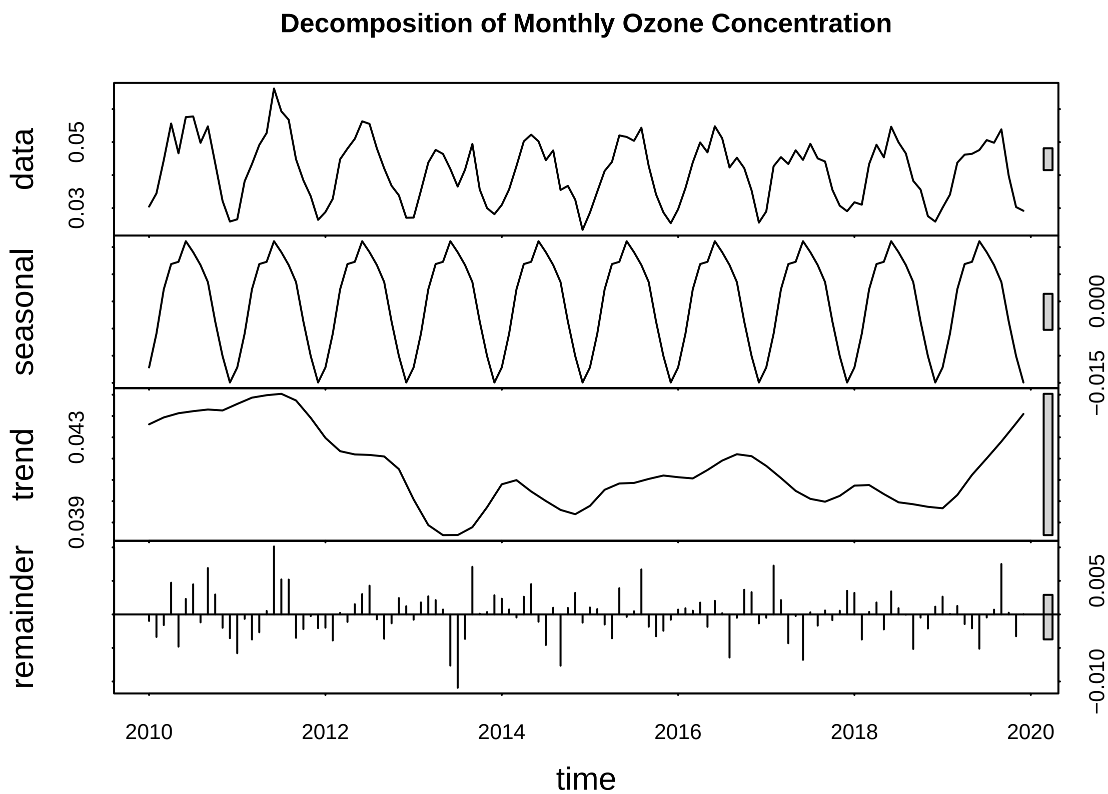
```r
#10
GaringerOzone.daily.ts <- ts(GaringerOzone$Daily.Max.8.hour.Ozone.Concentration,
                             start = c(2010, 1),
                             end = c(2019, 365),
                             frequency = 365)

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

#11.
# Daily decomposition
GaringerOzone.daily.decomp <- stl(GaringerOzone.daily.ts, s.window = "periodic")
plot(GaringerOzone.daily.decomp, main = "Decomposition of Daily Ozone Concentration")
```




```r
# Monthly decomposition
GaringerOzone.monthly.decomp <- stl(GaringerOzone.monthly.ts, s.window = "periodic")
plot(GaringerOzone.monthly.decomp, main = "Decomposition of Monthly Ozone Concentration")
```



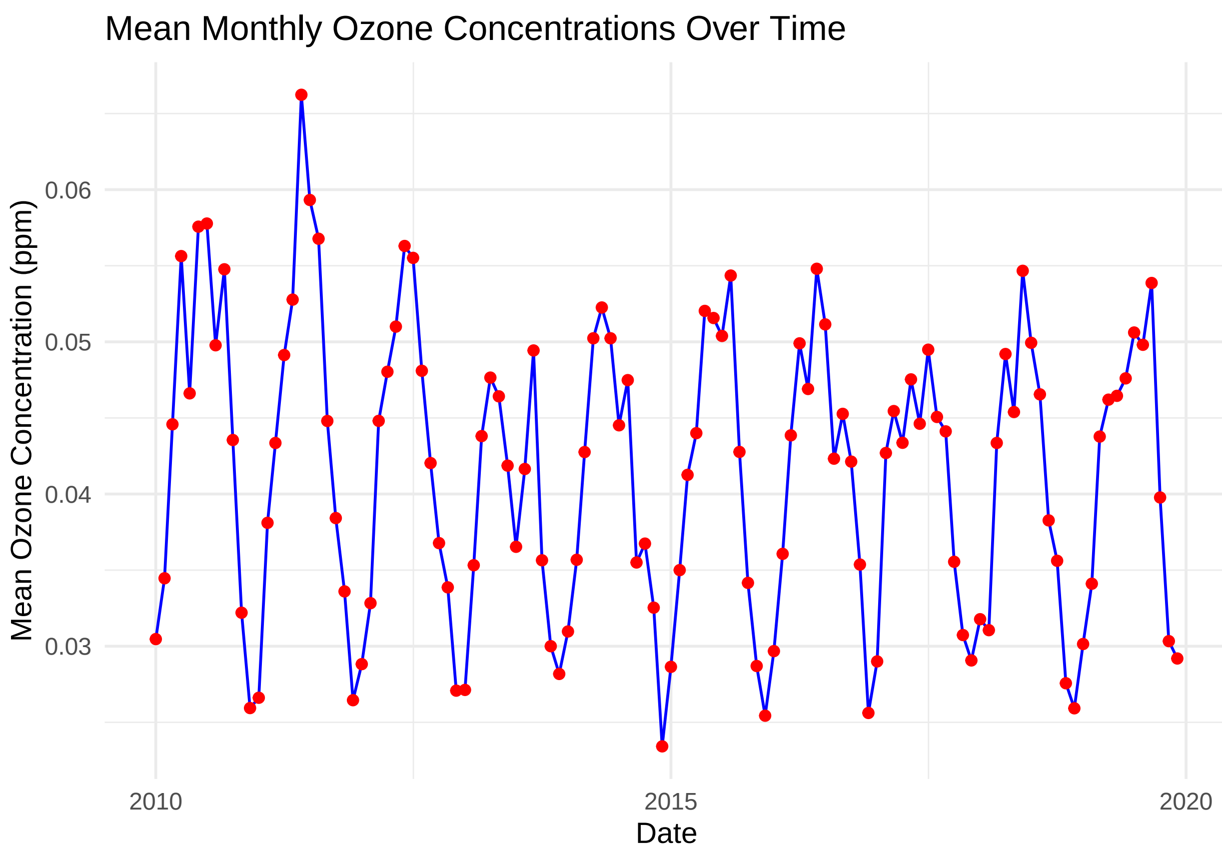
12. Run a monotonic trend analysis for the monthly Ozone series. In this case the seasonal Mann-Kendall is most appropriate; why is this?

```
#12
```

Answer:

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_line(color = "blue") +
  geom_point(color = "red", size = 1.5) +
  labs(title = "Mean Monthly Ozone Concentrations Over Time",
       x = "Date",
       y = "Mean Ozone Concentration (ppm)") +
  theme_minimal()
```



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: Mean monthly ozone concentrations at this station display a clear seasonal pattern, with recurring peaks and dips each year. While there is some variation from year to year, there doesn't seem to be a noticeable long-term increase or decrease in concentration over the decade.

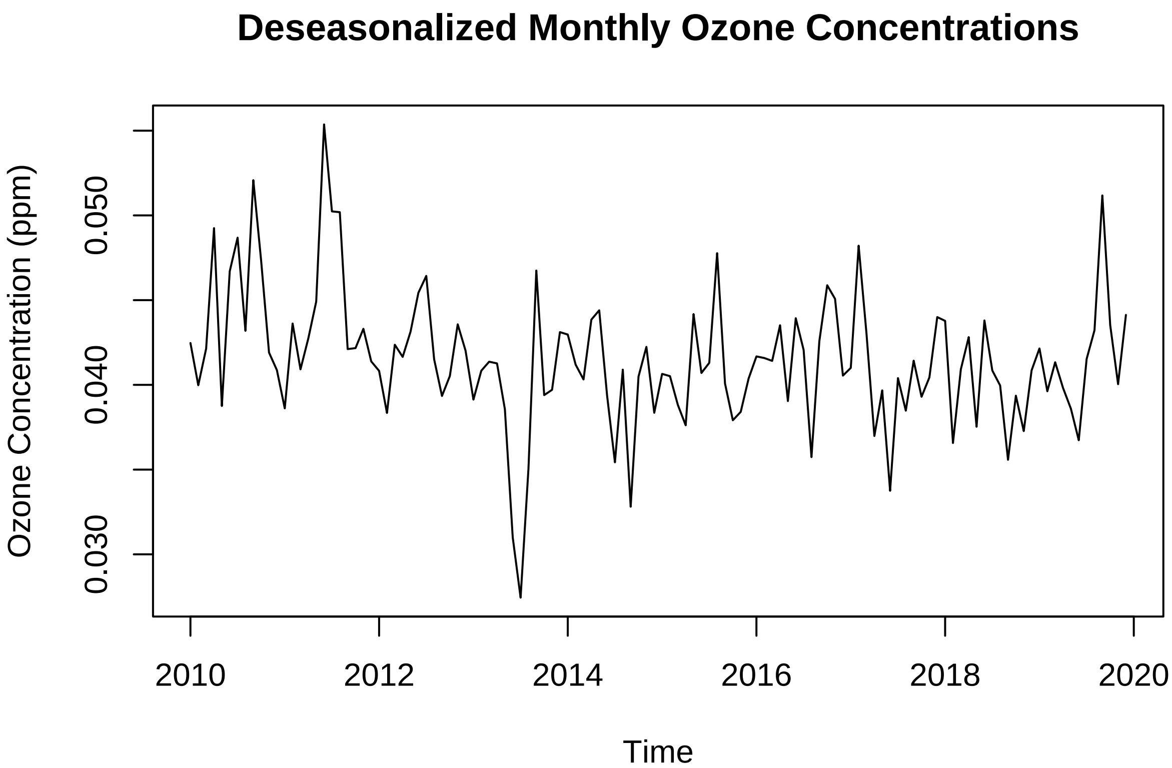
15. Subtract the seasonal component from the `GaringerOzone.monthly.ts`. Hint: Look at how we extracted the series components for the EnoDischarge on the lesson Rmd file.

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.decomp <- decompose(GaringerOzone.monthly.ts)

GaringerOzone.deseasonalized <- GaringerOzone.monthly.ts - GaringerOzone.monthly.decomp$seasonal

plot(GaringerOzone.deseasonalized,
     main = "Deseasonalized Monthly Ozone Concentrations",
     ylab = "Ozone Concentration (ppm)", xlab = "Time")
```


Deseasonalized Monthly Ozone Concentrations
Ozone Concentration (ppm)
0.030
0.040
0.050
2010
2012
2014
2016
2018
2020
Time

```
#16
```

Answer: