IreneChang_A08_TimeSeries.Rmd

Irene Chang

Spring 2023

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/EDA-Spring2023"
```

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

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.0
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.2.1     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
##
```

```
## Attaching package: 'lubridate'
##
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
##
##
##
## Attaching package: 'zoo'
##
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(trend); library(Kendall); library(tseries)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
mytheme <- theme_classic(base_size = 14) +
  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

EPA.Ozone.2010<- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2010_raw.csv")
EPA.Ozone.2011<- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2011_raw.csv")
EPA.Ozone.2012<- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2012_raw.csv")
EPA.Ozone.2013<- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2013_raw.csv")
EPA.Ozone.2014<- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2014_raw.csv")
EPA.Ozone.2015<- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2015_raw.csv")
EPA.Ozone.2016<- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2016_raw.csv")
EPA.Ozone.2017<- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2017_raw.csv")
EPA.Ozone.2018<- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2018_raw.csv")
EPA.Ozone.2019<- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2019_raw.csv")

GaringerOzone <- rbind(EPA.Ozone.2010, EPA.Ozone.2011, EPA.Ozone.2012, EPA.Ozone.2013, EPA.Ozone.2014, I
```

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)

#4

GaringerOzone.newcolumns <-
  select(GaringerOzone, Date, Daily.Max.8.hour.Ozone.Concentration, DAILY_AQI_VALUE)

#5

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

colnames(Days)[1] = "Date"

#6

GaringerOzone <-
  left_join(Days, GaringerOzone.newcolumns)
```

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

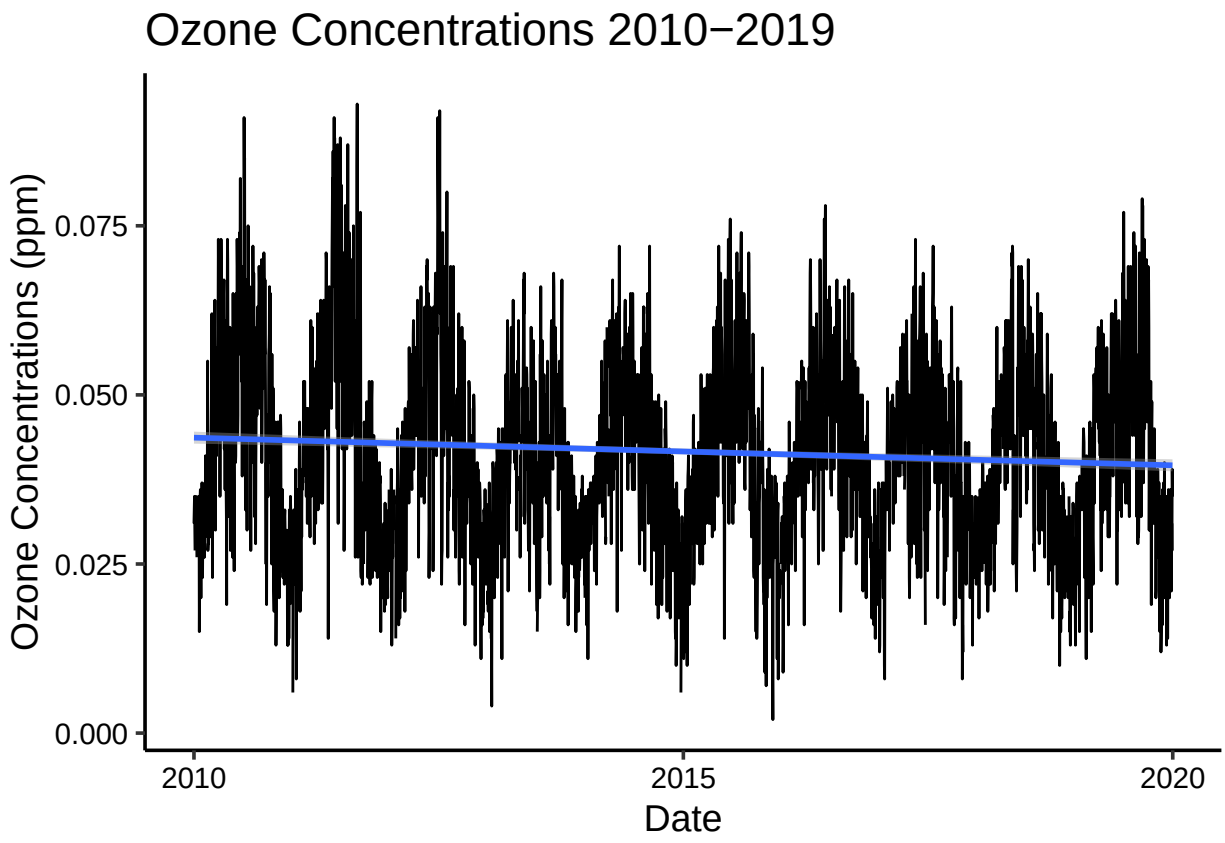
Visualize

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?

```
#7

GaringerOzone.lineplot <-
  ggplot(GaringerOzone,
         aes(x=Date, y=Daily.Max.8.hour.Ozone.Concentration)) +
  geom_line() +
  geom_smooth(method="lm") +
  labs(
    title = "Ozone Concentrations 2010-2019",
    x = "Date",
    y= "Ozone Concentrations (ppm)"
```

```
  )

print(GaringerOzone.lineplot)
```

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

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



Answer: Yes slightly, there is a slight downward trend in the ozone concentrations over time based on the line of best fit.

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

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
```

```
summary(GaringerOzone$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
```

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

summary(GaringerOzone.clean$OzoneConcentration.clean)
```

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

Answer: We used a linear interpolation instead of a piecewise constant or spline interpolation because the data has short periods of missing data (NAs). Piecewise constant would not be a good fit because from the data, it is clear that there is enough variation in the data that we should not assume that the missing data should be equal to the measurement made nearest to that date. We didn't use a spline interpolation because the data is not taken at irregular intervals and there doesn't appear to be any extreme outliers.

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)

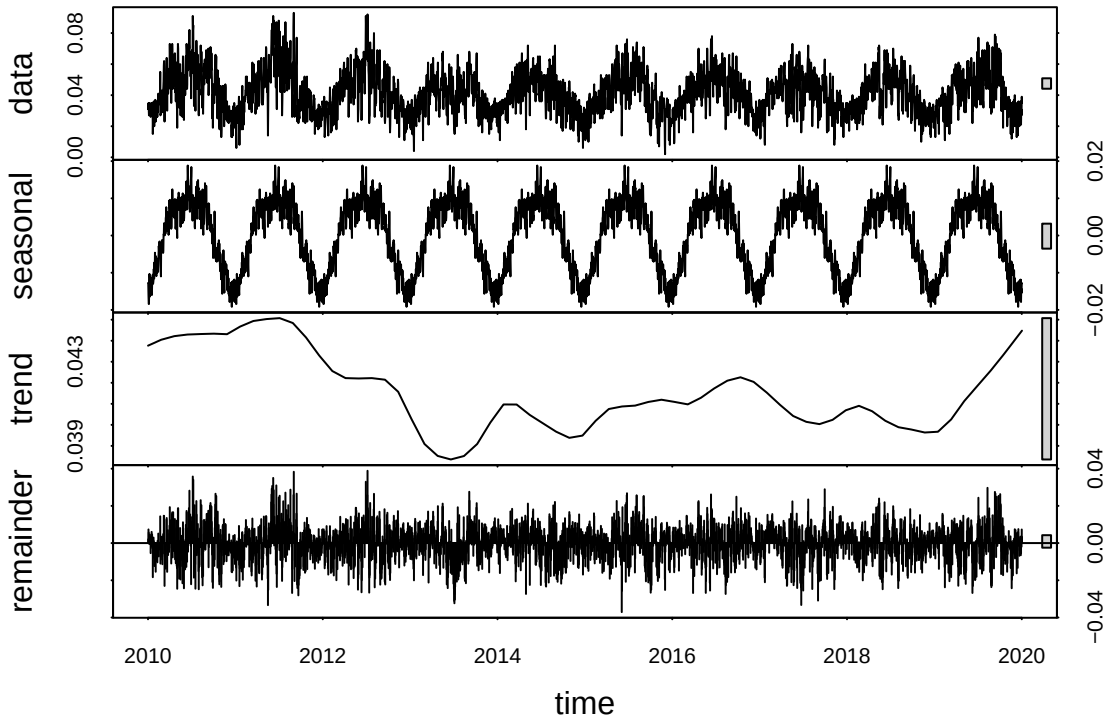
```
#9

GaringerOzone.newdatecolumn <-
  GaringerOzone.clean %>%
  mutate(Month= month(Date), Year=year(Date))

GaringerOzone.monthly <-
  GaringerOzone.newdatecolumn %>%
  mutate(Date= my(paste0(Month, "-", Year))) %>%
  group_by(Date) %>%
  summarise(MeanOzone= mean(OzoneConcentration.clean))
```

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.

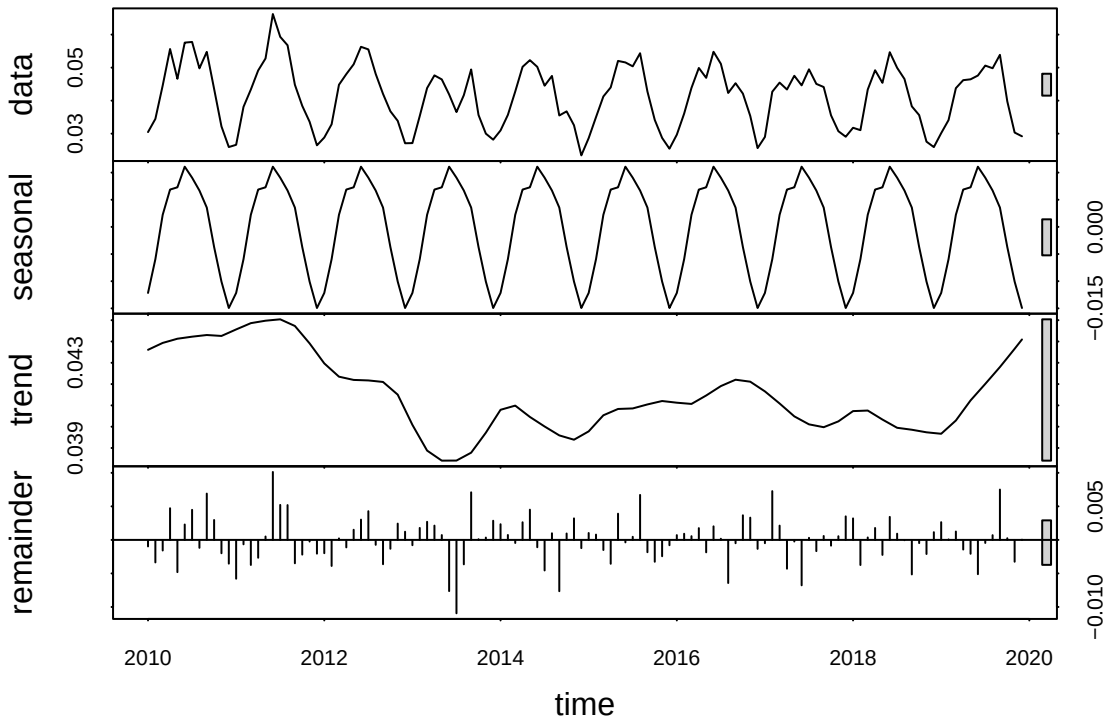
```r
#10

GaringerOzone.daily.ts <-
  ts(GaringerOzone.clean$OzoneConcentration.clean,
     start = c(2010,1,1),
     frequency = 365)

GaringerOzone.monthly.ts <-
  ts(GaringerOzone.monthly$MeanOzone,
     start = c(2010,1),
     end = c(2019, 12),
     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

GaringerOzone.monthly.smk <-
  Kendall::SeasonalMannKendall(GaringerOzone.monthly.ts)
summary(GaringerOzone.monthly.smk)
```

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

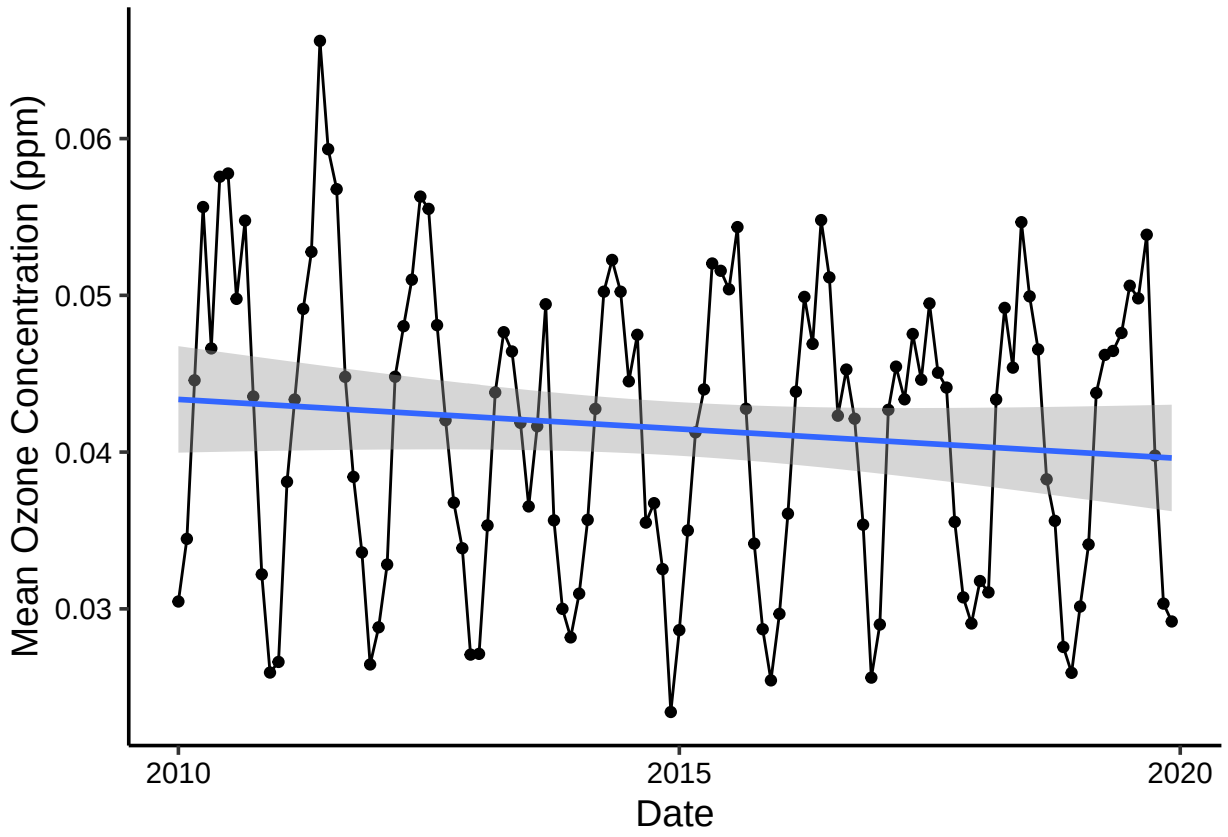
Answer: We use the seasonal Mann-Kendall because the ozone concentrations are affected by seasonality. We can see this because there is a pattern within a specific year but the pattern is repeated from year to year. The data is also non-parametric.

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

GaringerOzone.monthly.plot <-
  ggplot(GaringerOzone.monthly, aes(x = Date, y = MeanOzone)) +
```

```
  geom_point() +
  geom_line() +
  ylab("Mean Ozone Concentration (ppm)") +
  geom_smooth( method = lm )
print(GaringerOzone.monthly.plot)
```

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



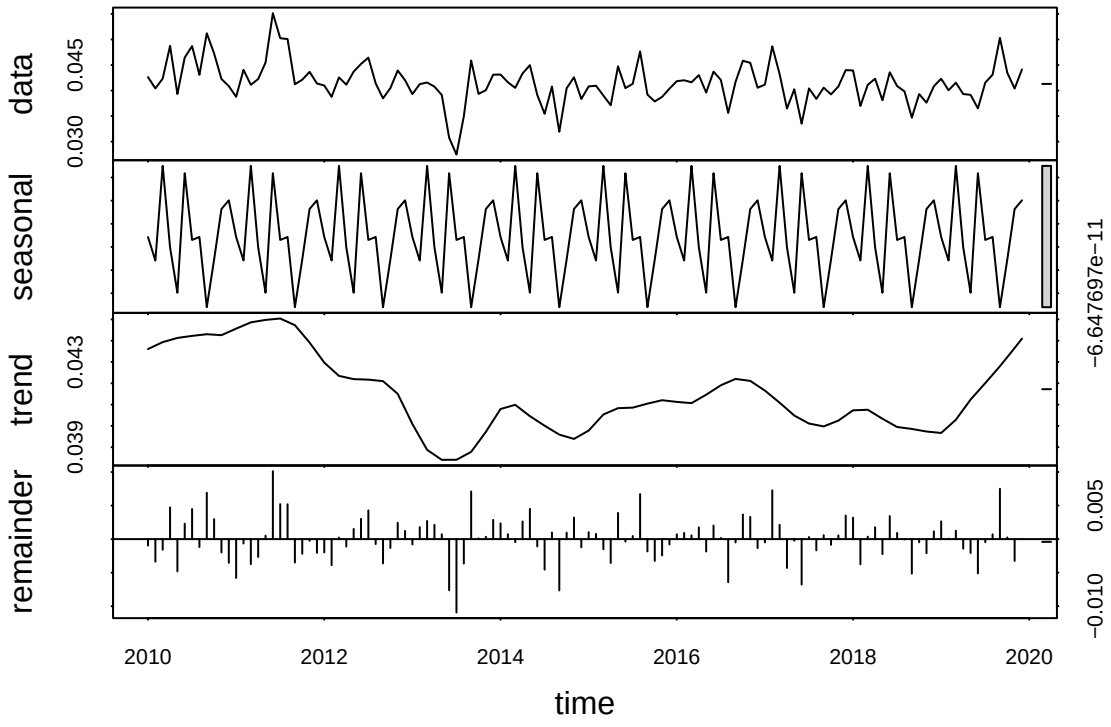
14. To accompany your graph, summarize your results in context of the research question. Include output from the statistical test in parentheses at the end of your sentence. Feel free to use multiple sentences in your interpretation.

    Answer: The statistical tests showed that there was a p-value of 0.0467 (slightly significant) and because of that we can conclude that there was a change on ozone concentration over the 2010s at this station. When looking at the graph, the slope of the line is negative, meaning that the concentrations on average has decreased over time. We do see an occurance of seasonality, so it would be in our best interest to remove seasonality to see if this trend holds true regardless of seasonality.

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

GaringerOzone.components <-
  as.data.frame(GaringerOzone.monthly_Decomposed$time.series[,1:3])

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

GaringerOzone.monthly.noseasonal = GaringerOzone.components[,4] - GaringerOzone.components[,1]

GaringerOzone.monthly.noseasonal.ts <-
  ts(GaringerOzone.monthly.noseasonal, start = c(2010,1), end = c(2019, 12), frequency = 12)

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




```r
#16

GaringerOzone.monthly.noseasonal.mk <-
  Kendall::MannKendall(GaringerOzone.monthly.noseasonal.ts)
summary(GaringerOzone.monthly.noseasonal.mk)
```

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

Answer: The p-value went from 0.0467 (slightly significant) on the seasonal Mann Kendall test to 0.00754 (significant) on the Mann Kendall test, when removing seasonality. Because the seasonality created varied trends within each year, it increased our p-value. When removing seasonality, we can see the true trend of ozone concentrations at this station during this time period (ozone concentrations declined). Therefore, we are confident that ozone concentrations in the 2010s at this station changed.