Assignment 7: Time Series Analysis

Changxin Yu

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
getwd()
```

```
## [1] "E:/things/Duke University/study/2022 Fall/ENVIRON 872/EDA-Fall2022/Assignments"
```

```
setwd("E:/things/Duke University/study/2022 Fall/ENVIRON 872/EDA-Fall2022")
library(tidyverse)
library(lubridate)
library(zoo)
library(trend)

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

#2
OzoneFiles = list.files(path = "./Data/Raw/Ozone_TimeSeries/",
                        pattern="*.csv", full.names=TRUE)
GaringerOzone <- OzoneFiles %>%
  plyr::ldply(read.csv, stringsAsFactors = TRUE)
```

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")

# 4
GaringerOzone <- 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"), "day"))
colnames(Days) <- "Date"

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

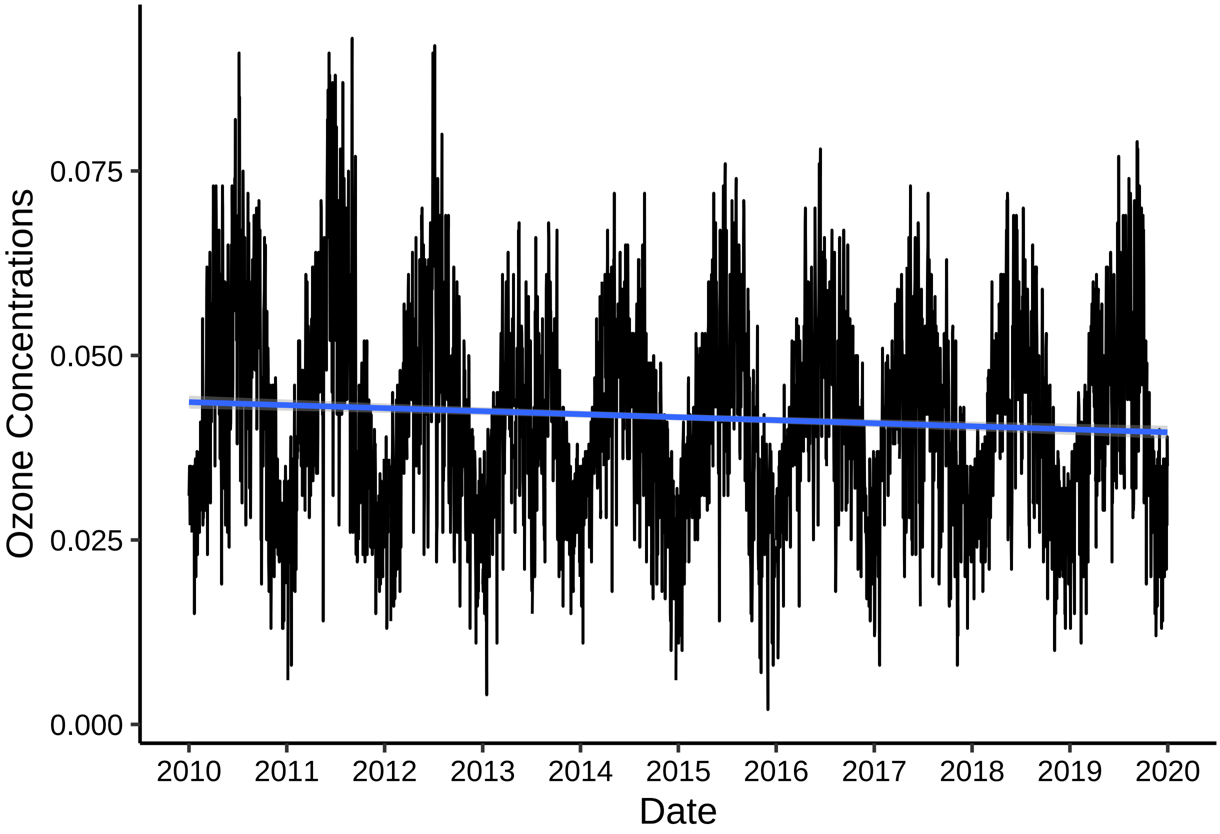
```
## Joining, by = "Date"
```

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
ggplot(GaringerOzone, aes(x = Date, y = Daily.Max.8.hour.Ozone.Concentration)) +
  geom_line() +
  labs(y = "Ozone Concentrations") +
  geom_smooth(method="lm")+
  scale_x_date(date_breaks = "1 years", date_labels = "%Y")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



Answer: From the figure, the linear fitting line is nearly flat and has a little downward trend, which suggests that there might be no trend or a little 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?

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

Answer: In this case, piecewise constant interpolation might be less accurate than linear interpolation because the value of ozone concentrations has been changing over time. And since there are few missing values, it is not necessary to use spline interpolation. Linear interpolation can do a good job with small computations.

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 %>%
  mutate(Month = month(GaringerOzone$Date)) %>%
  mutate(Year = year(GaringerOzone$Date)) %>%
  group_by(Year, Month) %>%
  summarise(mean_ozoneCon = mean(Daily.Max.8.hour.Ozone.Concentration))
```

```
## ‘summarise()‘ has grouped output by ’Year’. You can override using the
## ‘.groups‘ argument.
```

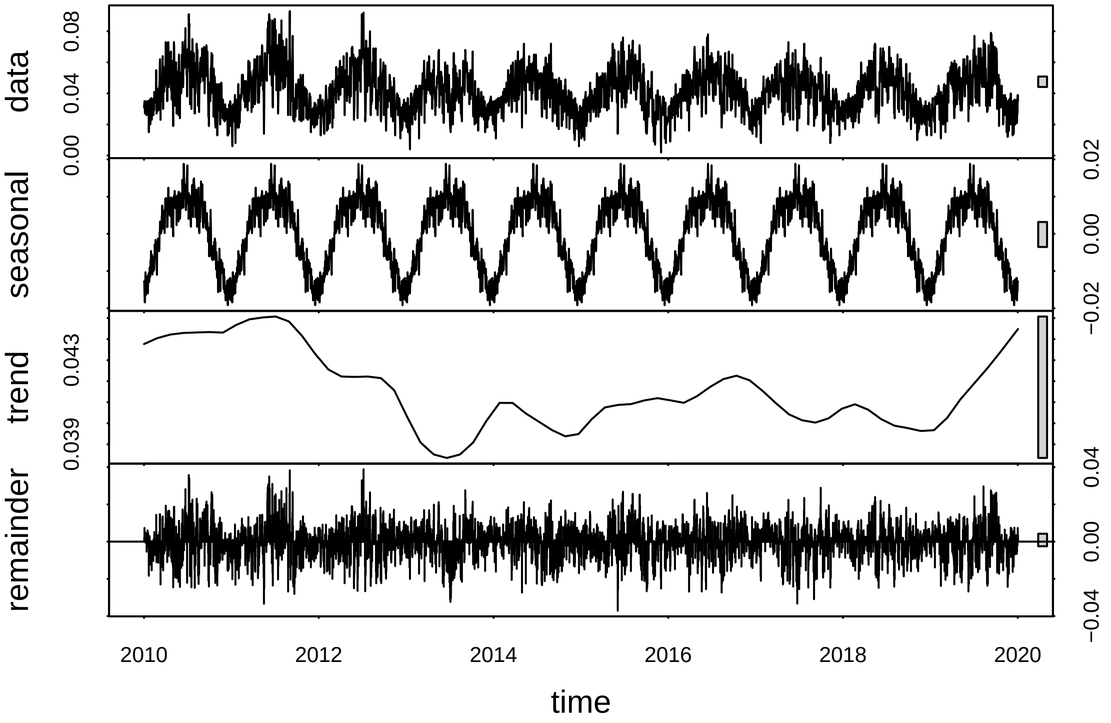
```
GaringerOzone.monthly$Date <- seq(as.Date("2010-01-01"),
                                  as.Date("2019-12-31"), "month")
```

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.

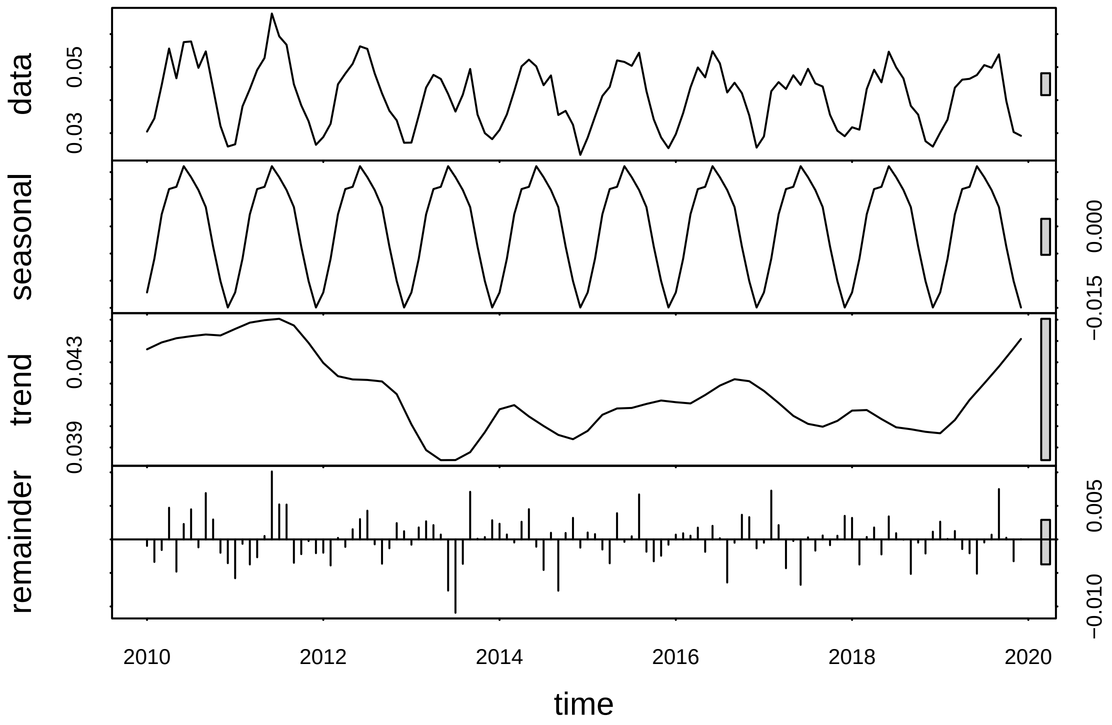
```
#10
GaringerOzone.daily.ts <-
  ts(GaringerOzone$Daily.Max.8.hour.Ozone.Concentration,
     start = c(2010,1,1), frequency = 365)
GaringerOzone.monthly.ts <- ts(GaringerOzone.monthly$mean_ozoneCon,
                             start = c(2010,1), frequency = 12)
```

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

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



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

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

```
summary(Ozone_monthly_trend)
```

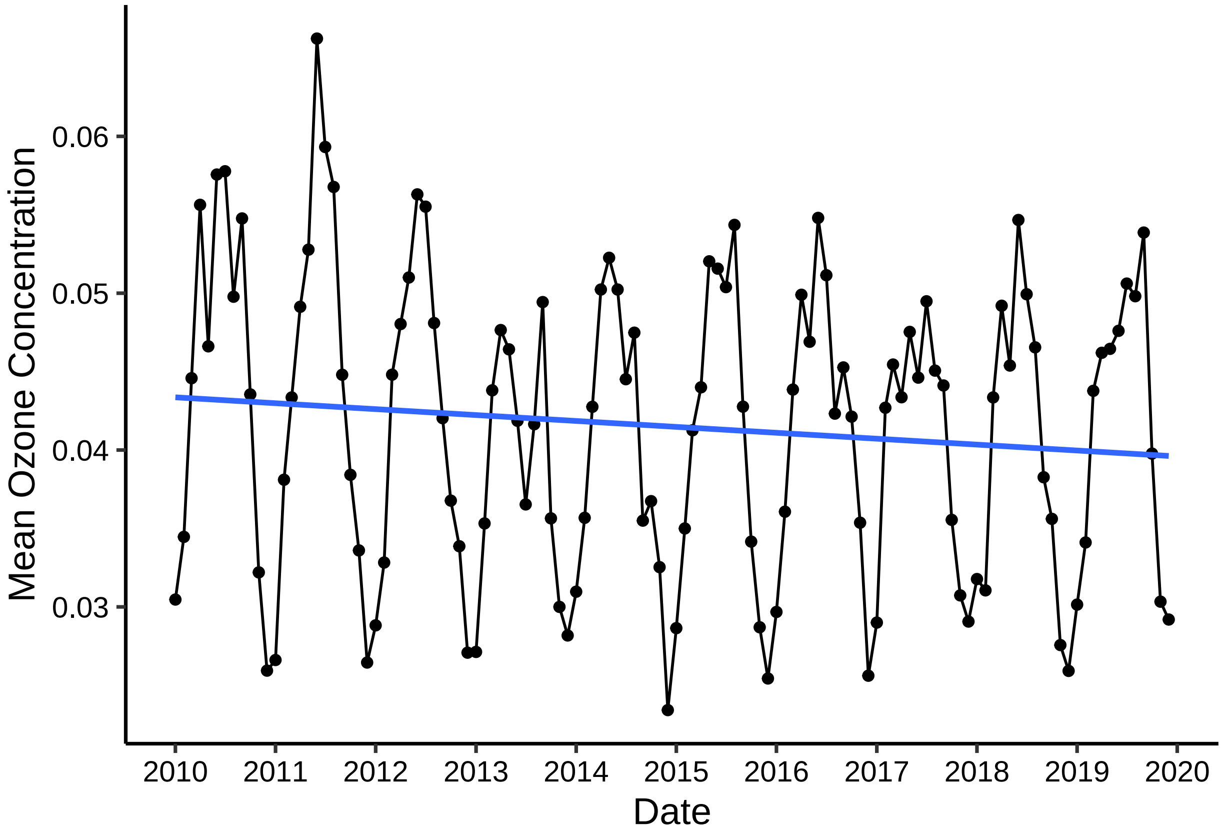
```
## Score =  -77 , Var(Score) = 1499
## denominator =  539.4972
## tau = -0.143, 2-sided pvalue =0.046724
```

Answer: Because the figure of decomposition of monthly time series object showed that it has a strong seasonal component, which means that the ozone concentration has a seasonal cycle, we need to use seasonal Mann-Kendall in this case. Other monotonic trend tests we have learned in this class could not deal with seasonality.

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.

```r
# 13
ggplot(GaringerOzone.monthly, aes(x = Date, y = mean_ozoneCon)) +
  geom_point() +
  geom_line() +
  labs(y = "Mean Ozone Concentration") +
  geom_smooth(method="lm", se=FALSE) +
  scale_x_date(date_breaks = "1 years", date_labels = "%Y")
```

```
## '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: From the figure, there is a little descending trend of mean monthly ozone concentrations. The null hypothesis of seasonal Mann-Kendall test is the data is stationary over time. From SMK above, the p-value is smaller than 0.05, so we reject the null hypothesis, which means that ozone concentrations have changed over the 2010s at this station (p-value = 0.046724).

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.adjusted.ts <-
  GaringerOzone.monthly.ts -
  GaringerOzone.monthly.Decomposed$time.series[,1]

#16
Ozone_monthly_trend2 <- Kendall::MannKendall(GaringerOzone.monthly.adjusted.ts)
Ozone_monthly_trend2
```

```
## tau = -0.165, 2-sided pvalue =0.0075402
```

```
summary(Ozone_monthly_trend2)
```

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

> Answer: From Mann Kendall test, we still reject the null hypothesis since the p-value is smaller than 0.05. However, it seems that ozone concentrations have changed over the 2010s more significantly according to Mann Kendall test rather than seasonal Mann-Kendall test, because p-value here is 0.0075402, which is much more smaller than 0.046724.