Assignment 7: Time Series Analysis

Isaac Benaka

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] "/Users/isaacbenaka/Desktop/Fall 2022/872 - Data Analytics/EDA-Fall2022"
```

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

```
iketheme <- theme_bw(base_size = 8) + theme(axis.text = element_text(color = "black"),
    legend.position = "top")
theme_set(iketheme)

O10 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2010_raw.csv",
    stringsAsFactors = T)
O11 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2011_raw.csv",
    stringsAsFactors = T)
O12 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2012_raw.csv",
    stringsAsFactors = T)
O13 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2013_raw.csv",
    stringsAsFactors = T)
O14 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2014_raw.csv",
    stringsAsFactors = T)
O15 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2015_raw.csv",
    stringsAsFactors = T)
O16 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2016_raw.csv",
    stringsAsFactors = T)
O17 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2017_raw.csv",
    stringsAsFactors = T)
O18 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2018_raw.csv",
    stringsAsFactors = T)
O19 <- read.csv("./Data/Raw/Ozone_TimeSeries/EPAair_O3_GaringerNC2019_raw.csv",
    stringsAsFactors = T)

GaringerOzone <- bind_rows(O10, O11, O12,
    O13, O14, O15, O16, O17, O18, O19)
```

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"
```

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

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

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

```r
class(GaringerOzone$Date)
```

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

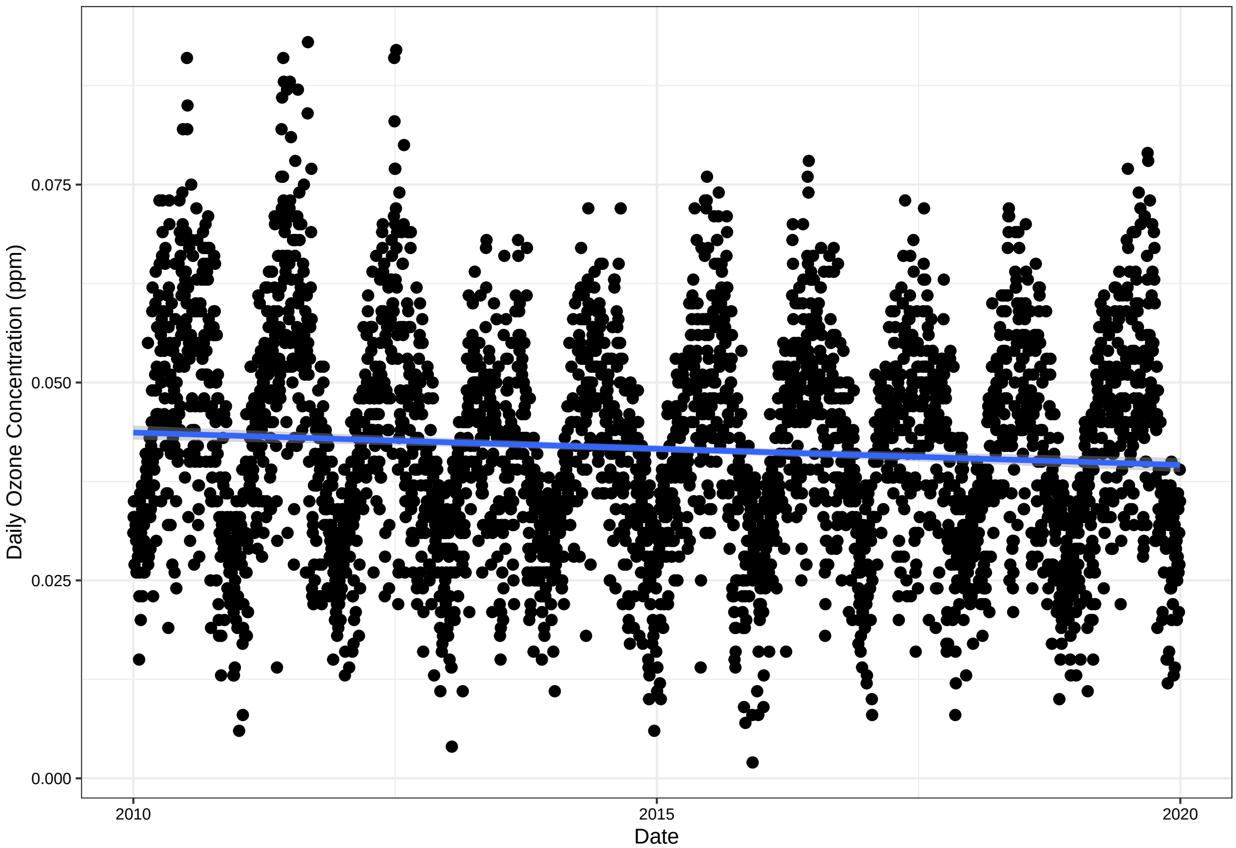
```r
# 6
GaringerOzone <- left_join(Days, GaringerOzone,
    by = "Date")
```

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
ggplot(GaringerOzone, (aes(x = Date, y = Daily.Max.8.hour.Ozone.Concentration))) +
    geom_point() + ylab("Daily Ozone Concentration (ppm)") +
    geom_smooth(method = "lm")
```

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

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

## Warning: Removed 63 rows containing missing values (geom_point).
```



Answer: The trend line appears to show that ozone concentrations are decreasing 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.clean <- GaringerOzone %>%
    mutate(Ozone.clean = zoo::na.approx(Daily.Max.8.hour.Ozone.Concentration))
```

Answer: We chose not to use piecewise interpolation because it copies data from nearby cells which would not work well with the fluctuating data we are dealing with. We chose not to use spline interpolation because it uses a quadratic function to replace NAs which would not result in a good fit to our data.

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.clean %>%
    mutate(Month = (month(Date))) %>%
    mutate(Year = (year(Date))) %>%
    group_by(Year, Month) %>%
    summarise(mean.Ozone.Conc = mean(Ozone.clean))
```

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

```r
GaringerOzone.monthly.plot <- GaringerOzone.clean %>%
    mutate(Month = (month(Date))) %>%
    mutate(Year = (year(Date))) %>%
    mutate(Date = my(paste0(Month, "-", Year))) %>%
    group_by(Date, Year, Month) %>%
    summarise(mean.Ozone.Conc = mean(Ozone.clean))
```

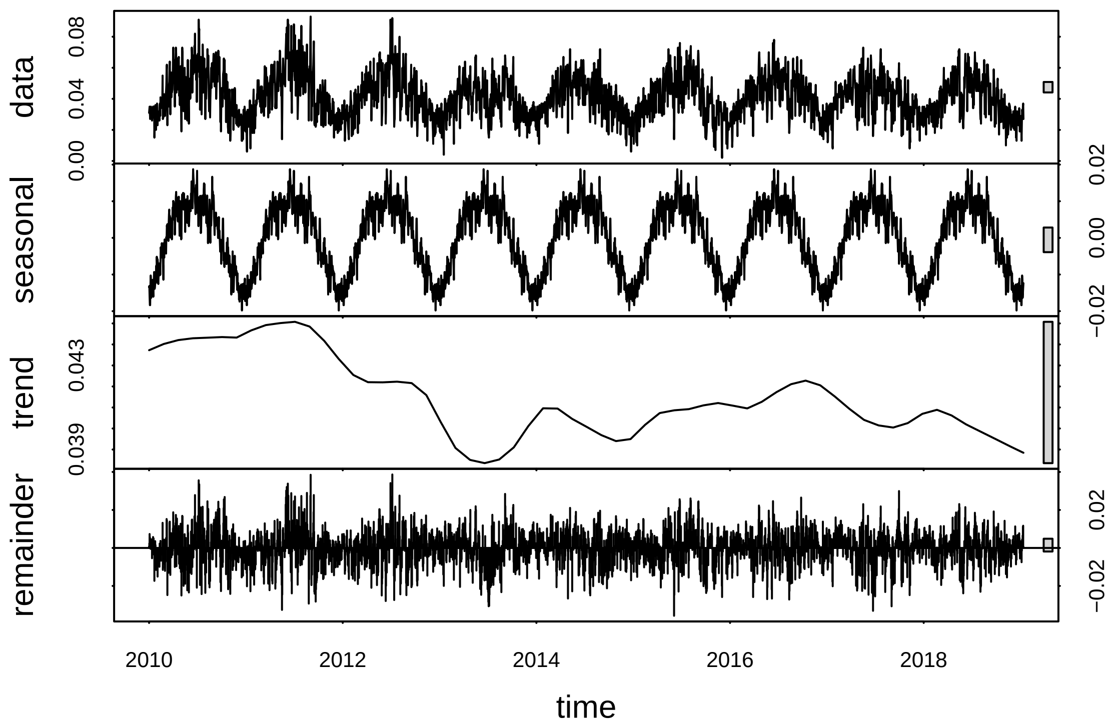
```
## `summarise()` has grouped output by 'Date', 'Year'. You can override using the
## `.groups` argument.
```

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.

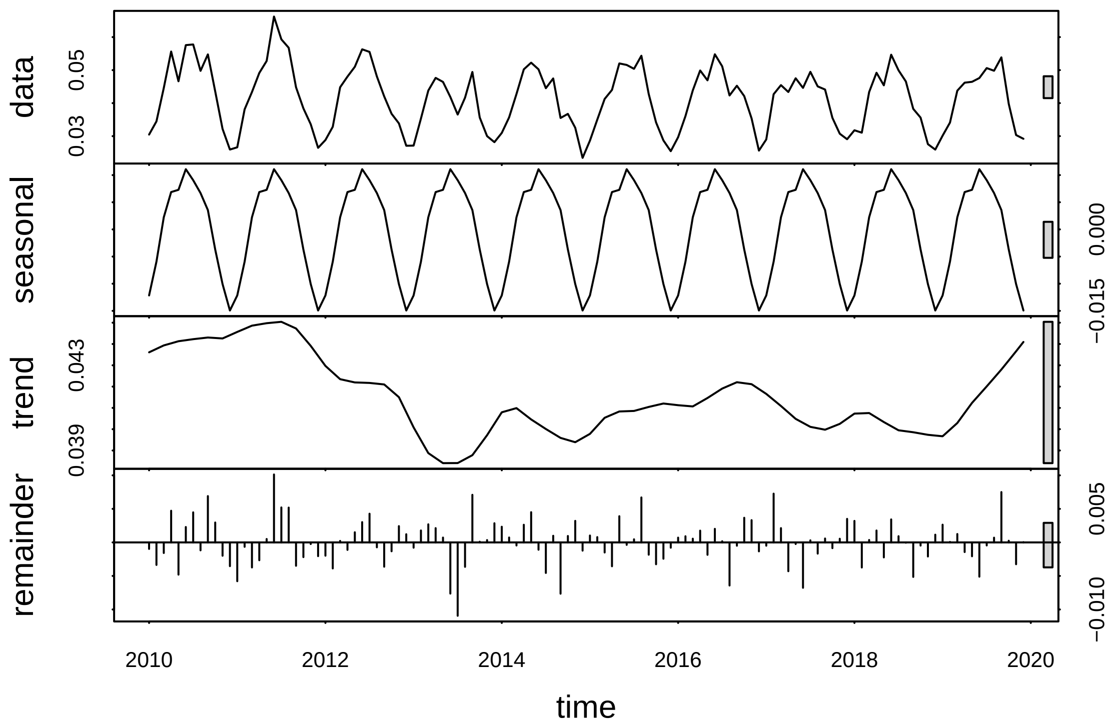
```r
# 10
f_day <- day(first(GaringerOzone.clean$Date))
f_month <- month(first(GaringerOzone.clean$Date))
f_year <- year(first(GaringerOzone.clean$Date))
l_day <- day(last(GaringerOzone.clean$Date))
l_month <- month(last(GaringerOzone.clean$Date))
l_year <- year(last(GaringerOzone.clean$Date))
GaringerOzone.daily.ts <- ts(GaringerOzone.clean$Ozone.clean,
    start = c(f_year, f_month, f_day), end = c(l_year,
        l_month, l_day), frequency = 365)

GaringerOzone.monthly.ts <- ts(GaringerOzone.monthly$mean.Ozone.Conc,
    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
Daily.decomp <- stl(GaringerOzone.daily.ts,
    s.window = "periodic")
plot(Daily.decomp)
```




```
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
Kendall::SeasonalMannKendall(GaringerOzone.daily.ts)
```

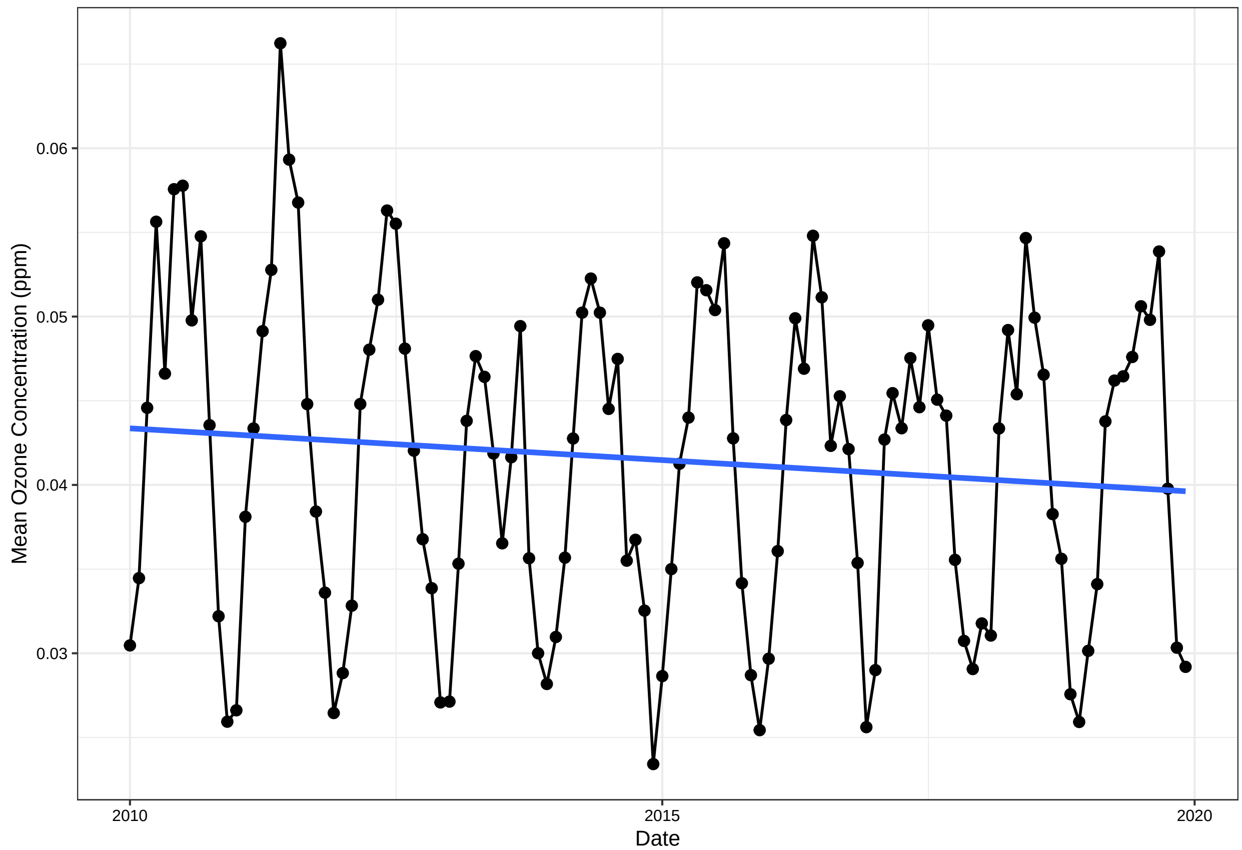
```
## tau = -0.0638, 2-sided pvalue =5.3605e-06
```

Answer: The seasonal Mann-Kendall function tests whether a trend is monotonic, or if it has a consistent upwards or downwards trend. The seasonal version of this test is best to employ in this situation because there are clear fluctuations every year given the season. The data moves up and down throughout the year because of seasonal changes, so we need a test that can account for and mitigate these changes to discern a trend. Given that the p-value of the test is <0.05, we accept the alternative hypotesis that a trend exists within the data.

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.plot, aes(x = Date,
    y = mean.Ozone.Conc)) + geom_point() +
    geom_line() + geom_smooth(method = "lm",
    se = FALSE) + ylab("Mean Ozone Concentration (ppm)")
```

```
## '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 data shows decreasing ozone concentrations through the 2010s at this station (tau = -0.0638, 2-sided pvalue =5.3605e-06).

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
Monthly.no.season <- Monthly.decomp$time.series[,
    2:3]

# 16
Kendall::MannKendall(Monthly.no.season)
```

```
## tau = -0.568, 2-sided pvalue =< 2.22e-16
```

Answer: By bemoving the seasonal component from the time series data, our p-value decreased significantly. Because the p-value became even smaller tthan it was before, the removal of the seasonal component accentuated the trend present in the data.