Assignment 5: Data Visualization

Soumya Mathew

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Visualization

Directions

1. Rename this file `<FirstLast>_A02_CodingBasics.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.

The completed exercise is due on Friday, Oct 14th @ 5:00pm.

Set up your session

1. Set up your session. Verify your working directory and load the tidyverse, lubridate, & cowplot packages. Upload the NTL-LTER processed data files for
nutrients and chemistry/physics for Peter and Paul Lakes (use the tidy [`NTL-LTER_Lake_Chemistry_Nutrients_PeterPa`
version) and the processed data file for the Niwot Ridge litter dataset (use the [`NEON_NIWO_Litter_mass_trap_Processe`
version).

2. Make sure R is reading dates as date format; if not change the format to date.

```
# 1
getwd()
```

```
## [1] "C:/Users/user/Desktop/Soumya/Year 2/EDA/EDA-Fall2022/Assignments"
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
# install.packages('cowplot')
library(cowplot)
```

```
##
## Attaching package: 'cowplot'
##
## The following object is masked from 'package:lubridate':
##
##     stamp
```

```
# dataloading
lakedata <- read.csv("C:/Users/user/Desktop/Soumya/Year 2/EDA/EDA-Fall2022/Data/Processed/NTL-LTER_Lake
    stringsAsFactors = TRUE)

ridgedata <- read.csv("C:/Users/user/Desktop/Soumya/Year 2/EDA/EDA-Fall2022/Data/Processed/NEON_NIWO_Li
    stringsAsFactors = TRUE)

# 2 changing datw format
# head(lakedata$sampledate, 5)
lakedata$sampledate <- c(ymd(lakedata$sampledate))
ridgedata$collectDate <- c(ymd(ridgedata$collectDate))
```

Define your theme

3. Build a theme and set it as your default theme.

```
# 3
library(tidyverse)
# setting theme
mytheme <- theme_classic(base_size = 8) +
    theme(axis.text = element_text(color = "black"),
        legend.position = "top")

theme_set(mytheme)
```

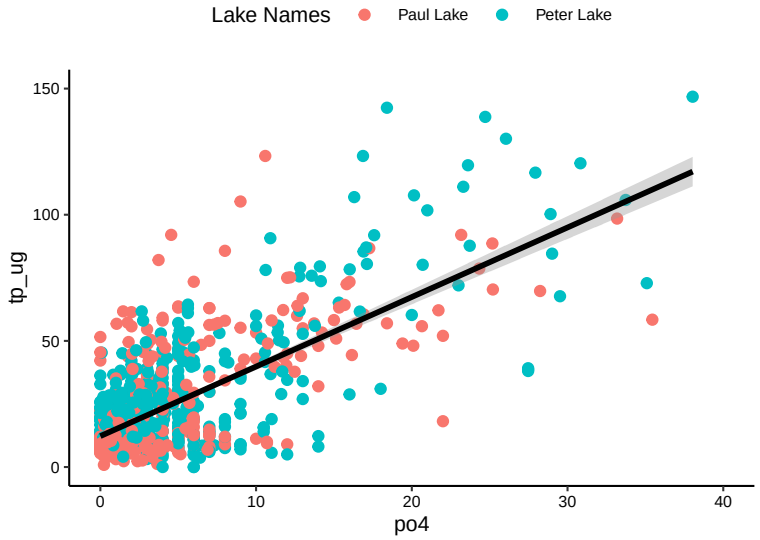
Create graphs

For numbers 4-7, create ggplot graphs and adjust aesthetics to follow best practices for data visualization. Ensure your theme, color palettes, axes, and additional aesthetics are edited accordingly.

4. [NTL-LTER] Plot total phosphorus (`tp_ug`) by phosphate (`po4`), with separate aesthetics for Peter and Paul lakes. Add a line of best fit and color it black. Adjust your axes to hide extreme values (hint: change the limits using `xlim()` and/or `ylim()`).

```
# 4
plot04 <- ggplot(lakedata, aes(x = po4, y = tp_ug,
    color = lakename)) + geom_point() + xlim(0,
    40) + ylim(0, 150) + labs(color = "Lake Names") +
    geom_smooth(method = lm, color = "black")
print(plot04)
```

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

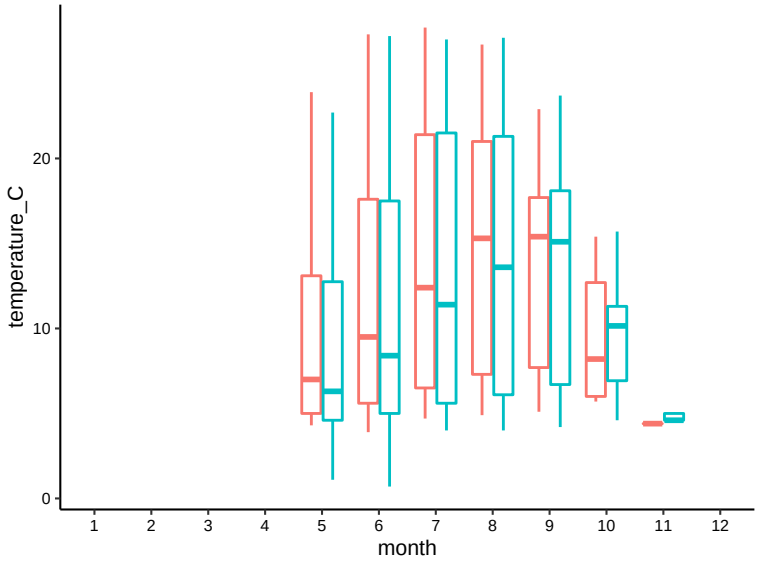


5. [NTL-LTER] Make three separate boxplots of (a) temperature, (b) TP, and

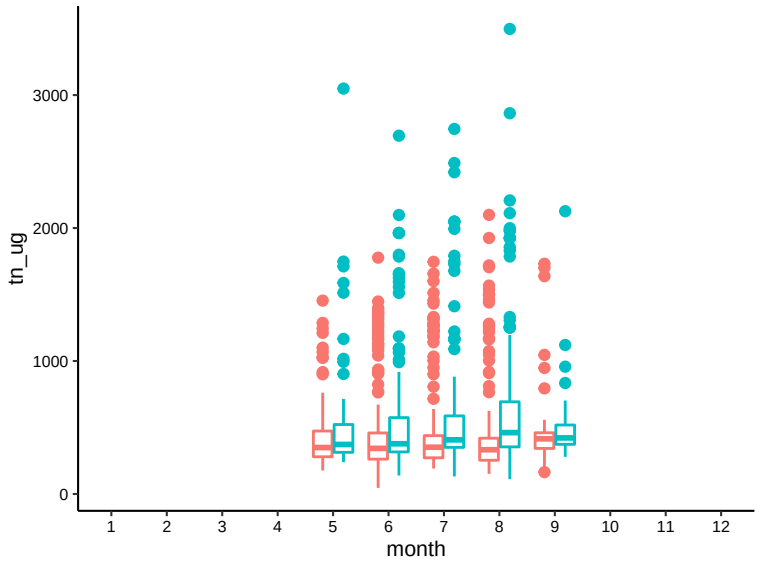
(c) TN, with month as the x axis and lake as a color aesthetic. Then, create a cowplot that combines the three graphs. Make sure that only one legend is present and that graph axes are aligned.

Tip: R has a build in variable called `month.abb` that returns a list of months; see https://r-lang.com/month-abb-in-r-with-example

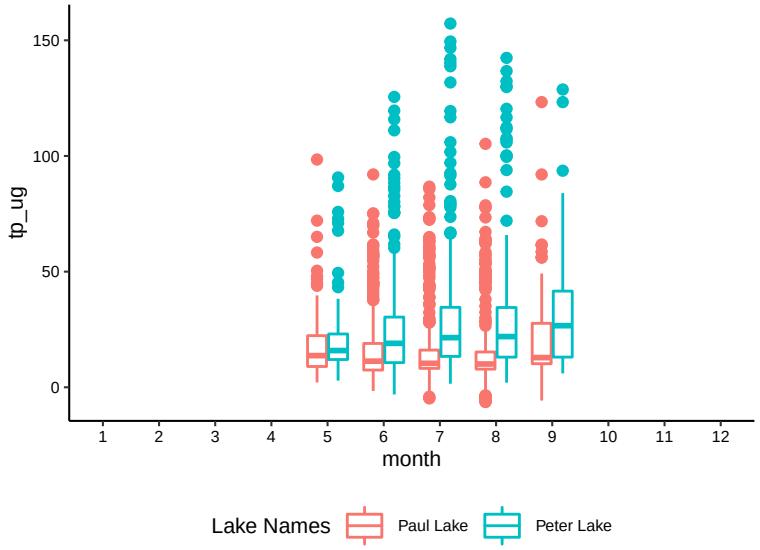
```
# 5 changing month into intergers
lakedata$month <- factor(lakedata$month,
    levels = c(1:12))

# temperature across months
plot05a <- ggplot(lakedata, aes(x = month,
    y = temperature_C)) + geom_boxplot(aes(color = lakename)) +
    theme(legend.position = "none") + scale_x_discrete(drop = FALSE)
print(plot05a)
```



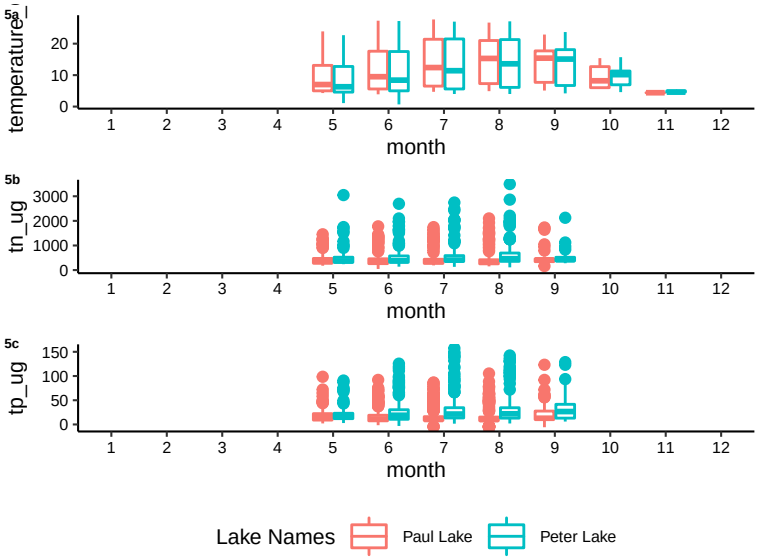

```
# nitrogen across months
plot05b <- ggplot(lakedata, aes(x = month,
    y = tn_ug)) + geom_boxplot(aes(color = lakename)) +
    theme(legend.position = "none") + scale_x_discrete(drop = FALSE)
print(plot05b)
```




```
# phosphate across months
plot05c <- ggplot(lakedata, aes(x = month,
    y = tp_ug)) + geom_boxplot(aes(color = lakename)) +
    theme(legend.position = "bottom") + labs(color = "Lake Names") +
    scale_x_discrete(drop = FALSE)
print(plot05c)
```




```
plot05 <- plot_grid(plot05a, plot05b, plot05c,
    nrow = 3, labels = c("5a", "5b", "5c"),
    label_size = 5, align = "v", rel_heights = c(10,
        10, 15))

print(plot05)
```



Question: What do you observe about the variables of interest over seasons and between lakes?

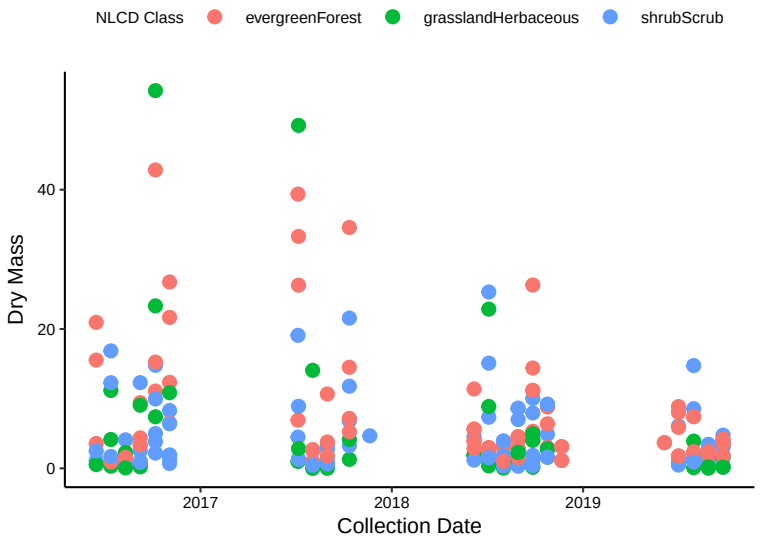
> Answer: The median of the temperature varied across months and remained higher for paul lake. The median level of phosphate across month have remained higher for peter lake than paul lake. Peter lake also have bigger (in terms of its unit of measurement) outliers than the paul lake.Median nitrogen level remained similar for the lake, however, slightly higher for peter lake. Peter lake also have higher variance than paul lake in terms of nitrogen level across months.

6. [Niwot Ridge] Plot a subset of the litter dataset by displaying only the "Needles" functional group.

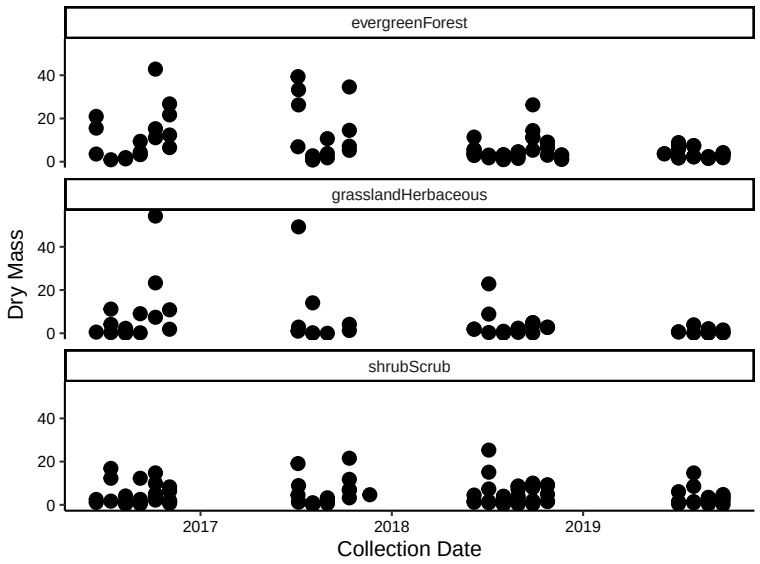
Plot the dry mass of needle litter by date and separate by NLCD class with a color aesthetic. (no need to adjust the name of each land use)

7. [Niwot Ridge] Now, plot the same plot but with NLCD classes separated into three facets rather than separated by color.

```
# 6
Plot06 <- ggplot(subset(ridgedata, functionalGroup ==
    "Needles"), aes(x = collectDate, y = dryMass,
    color = nlcdClass)) + geom_point(size = 2) +
    theme(legend.position = "top", legend.text = element_text(size = 6),
        legend.title = element_text(size = 6)) +
    ylab(expression("Dry Mass")) + xlab(expression("Collection Date")) +
    labs(color = "NLCD Class")

print(Plot06)
```




```
# 7
Plot07 <- ggplot(subset(ridgedata, functionalGroup ==
    "Needles"), aes(x = collectDate, y = dryMass)) +
    geom_point(size = 2) + theme(legend.position = "top",
    legend.text = element_text(size = 6),
    legend.title = element_text(size = 6)) +
    ylab(expression("Dry Mass")) + xlab(expression("Collection Date")) +
    labs(color = "NLCD Class") + facet_wrap(vars(nlcdClass),
    nrow = 3)

print(Plot07)
```



Question: Which of these plots (6 vs. 7) do you think is more effective, and why?

Answer: I prefer plot 7 over 6 as it shows changes in dry mass across time. Unlike plot 7, plot 6 is hard to read and no particualar trend can be discerned.