Assignment 5: Data Visualization

Changxin Yu

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A05_DataVisualization.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 21th @ 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] "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(cowplot)
NTL <- read.csv("./Data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv",stringsAsF
NEON <- read.csv("./Data/Processed/NEON_NIWO_Litter_mass_trap_Processed.csv",
                 stringsAsFactors = TRUE)

#2
NTL$sampledate <- as.Date(NTL$sampledate, format = "%Y-%m-%d")
NEON$collectDate <- as.Date(NEON$collectDate, format = "%Y-%m-%d")
```

Define your theme

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

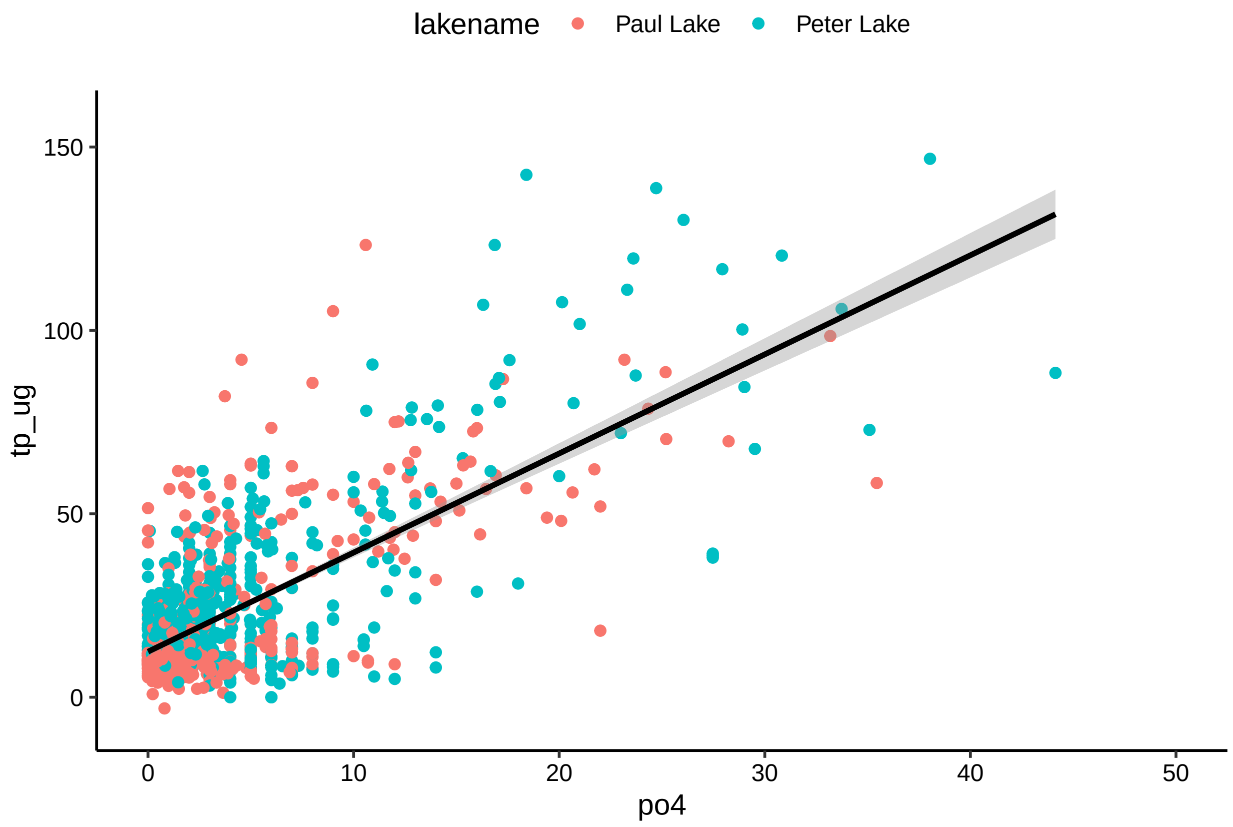
```
#3
mytheme <- theme_classic(base_size = 11) +
  theme(axis.text = element_text(color = "black"), legend.position = "top")
theme_set(mytheme)
```

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
ggplot(NTL, aes(x = po4, y = tp_ug, color = lakename))+
  geom_point()+
  geom_smooth(method="lm", color="black")+
  xlim(0, 50)
```

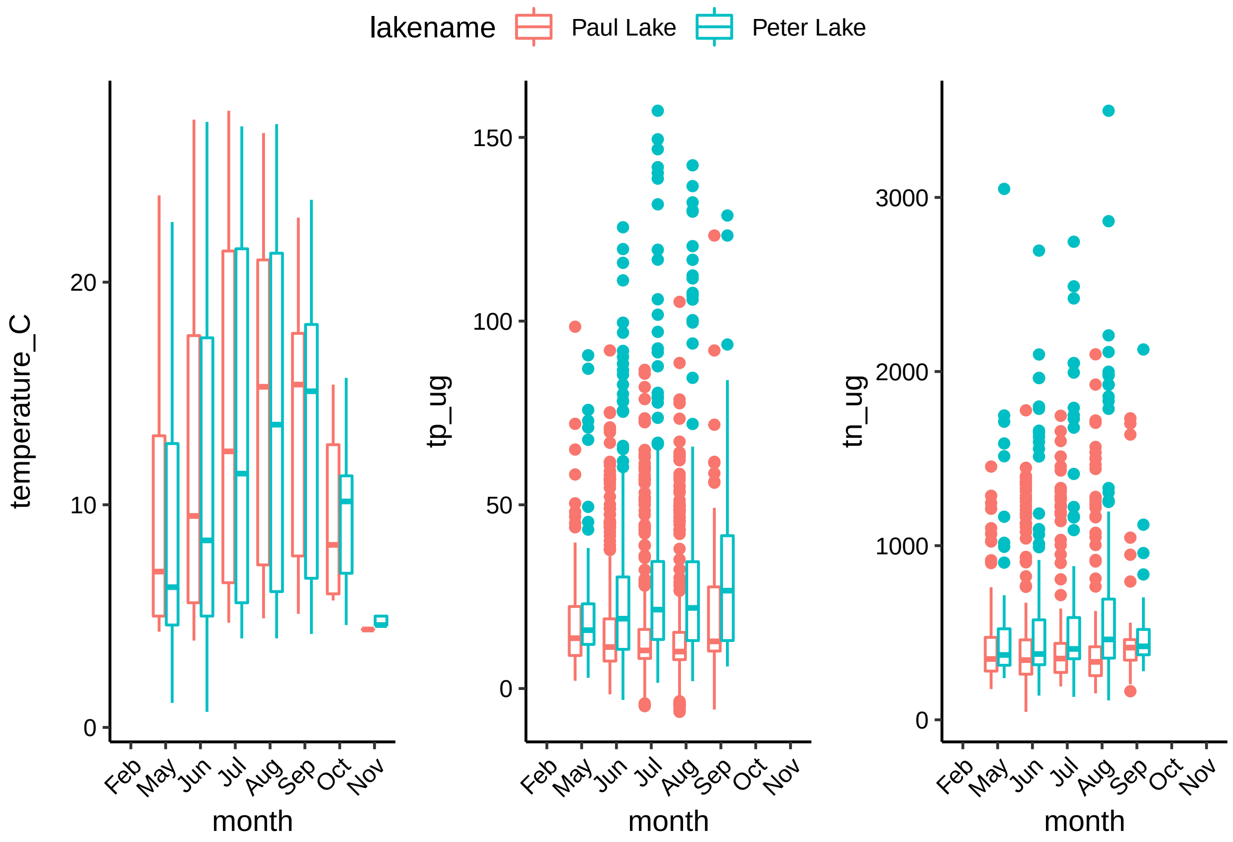


5. [NTL-LTER] Make three separate boxplots of (a) temperature, (b) TP, and

(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
NTL$month <- month.abb[NTL$month]
NTL$month <- factor(NTL$month, levels=month.abb)
tempMon <-ggplot(NTL, aes(x = month, y = temperature_C, color = lakename))+
  geom_boxplot()+
  theme(axis.text.x = element_text(angle = 45,  hjust = 1))
tpMon <-ggplot(NTL, aes(x = month, y = tp_ug, color = lakename))+
  geom_boxplot()+
  theme(axis.text.x = element_text(angle = 45,  hjust = 1))
tnMon <-ggplot(NTL, aes(x = month, y = tn_ug, color = lakename))+
  geom_boxplot()+
  theme(axis.text.x = element_text(angle = 45,  hjust = 1))
legend <- get_legend(tempMon)
prow <- plot_grid(tempMon+guides(color=F),
                  tpMon+guides(color=F),
                  tnMon+guides(color=F),
                  nrow = 1, align = 'vh', axis = 'tb')
plot_grid(legend, prow, nrow=2, rel_heights = c(0.1, 1))
```

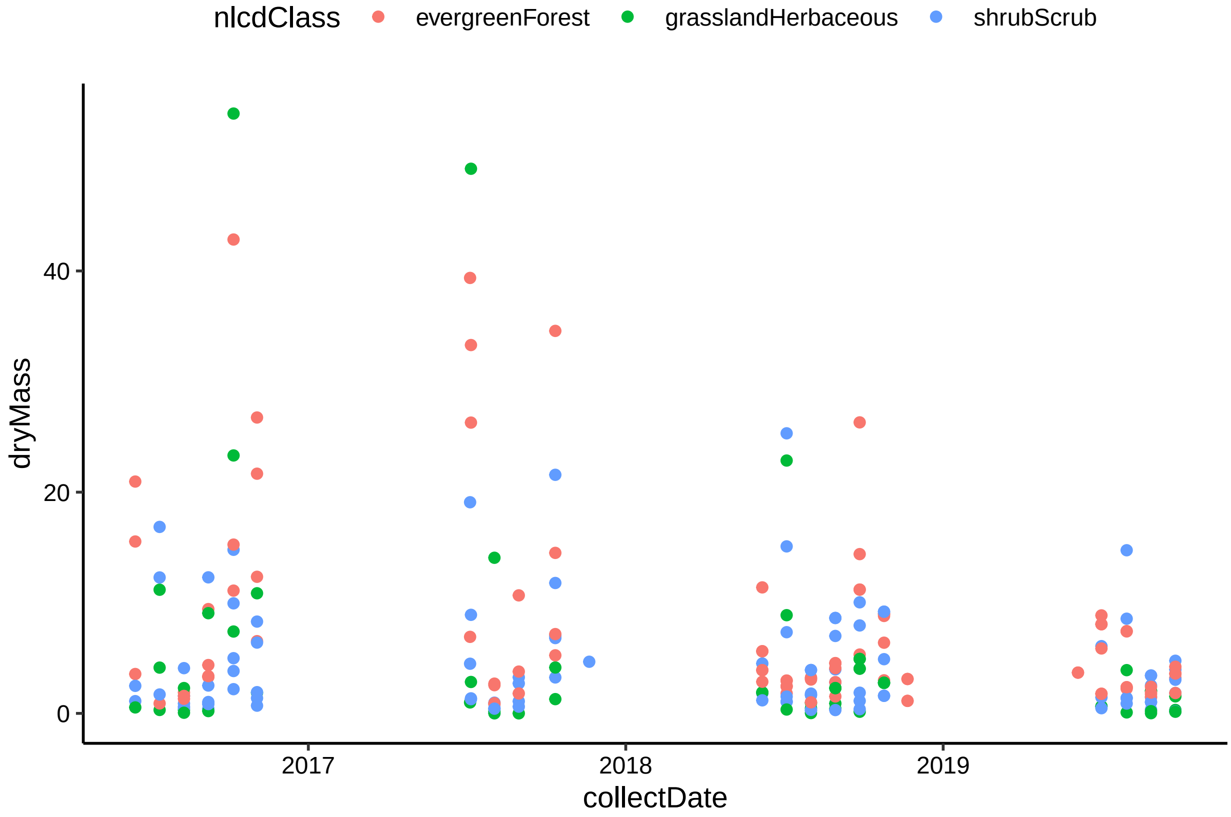


Question: What do you observe about the variables of interest over seasons and between lakes?

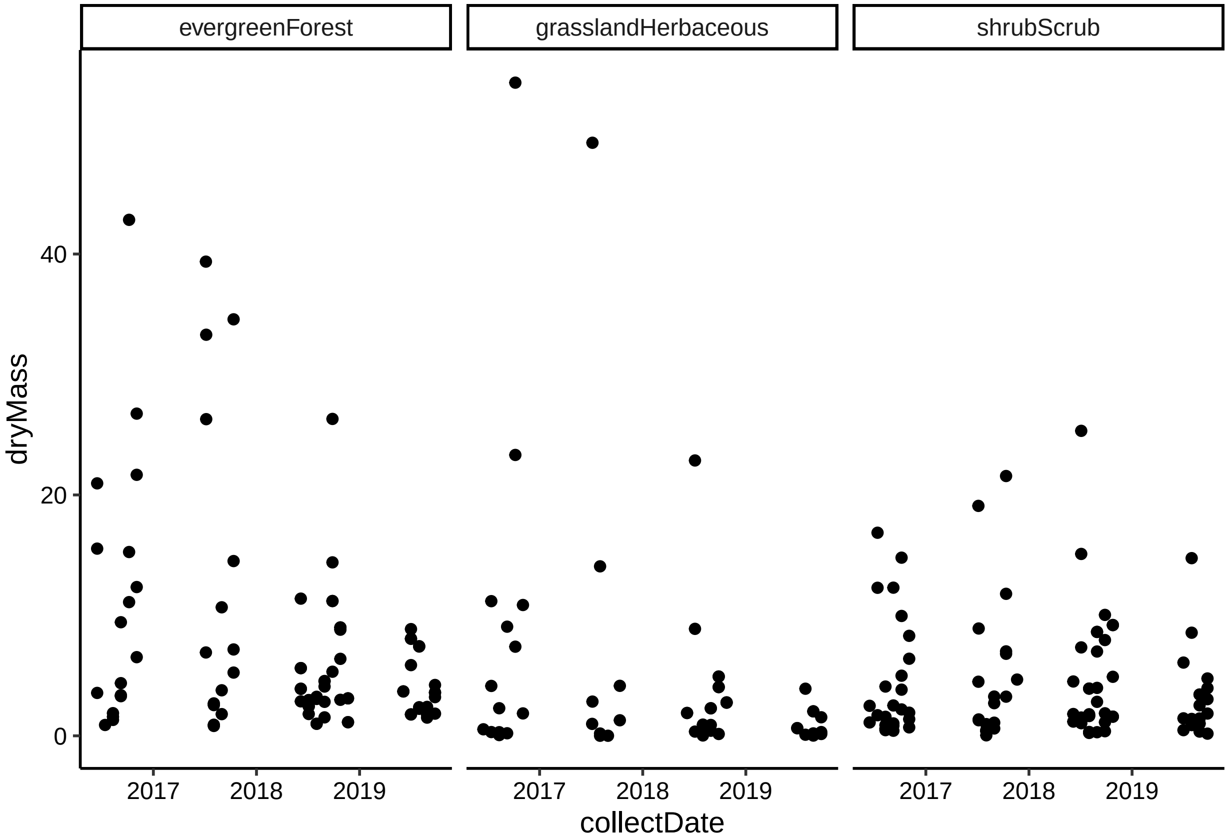
> Answer: From the plot, we can find that the temperature is relatively high in July and August and low in May, Oct and Nov. The temperature difference of two lakes is small. For the concentration of TP, there is not much different over seasons in Paul lake, while in Peter lake, it is getting higher from May to September. And the average TP concentration is higher in Peter lake than in Paul lake. For the concentration of TN, it is a little higher in Peter lake than in Paul lake. And in both two lakes, the differences over seasons are small.

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
ggplot(subset(NEON, functionalGroup=="Needles"),
       aes(x = collectDate, y = dryMass, color = nlcdClass))+
  geom_point()
```



```
#7
ggplot(subset(NEON, functionalGroup=="Needles"),
       aes(x = collectDate, y = dryMass))+
  geom_point()+
  facet_wrap(~nlcdClass)
```



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

> Answer: I think the plot from question 7 is more effective. In the previous plot (plot from question 6), it is hard to capture the information of different NLCD classes even if different colors are used. And the new plot is more intuitive by separating NLCD classes into three facets rather than gathering all points together.