Assignment 5: Data Visualization

Yixin Fang

Spring 2023

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 your code is tidy; use line breaks to ensure your code fits in the knitted output.
5. Be sure to **answer the questions** in this assignment document.
6. When you have completed the assignment, **Knit** the text and code into a single PDF file.

Set up your session

1. Set up your session. Load the tidyverse, lubridate, here & cowplot packages, and verify your home directory. 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_PeterPaul_Processed.csv` version) and the processed data file for the Niwot Ridge litter dataset (use the `NEON_NIWO_Litter_mass_trap_Processed.csv` version).

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

```
#1
getwd()
```

```
## [1] "D:/DKU/2023_Spring/ENV872/EDA-Spring2023"
```

```
library(tidyverse)
library(lubridate)
library(here)
library(cowplot)
library(ggpubr)
PeterPaul <- read.csv('./Data/Processed_KEY/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv')
Niwot <- read.csv('./Data/Processed_KEY/NEON_NIWO_Litter_mass_trap_Processed.csv')

#2
class(PeterPaul$sampledate)
```

```
## [1] "character"
```

```
PeterPaul$sampledate <- ymd(PeterPaul$sampledate)
class(Niwot$collectDate)
```

```
## [1] "character"
```

```
Niwot$collectDate <- ymd(Niwot$collectDate)
```

Define your theme

3. Build a theme and set it as your default theme. Customize the look of at least two of the following:

- Plot background
- Plot title
- Axis labels
- Axis ticks/gridlines
- Legend

```
#3
mytheme <- theme_classic(base_size = 12) +
  theme(plot.background = element_rect(color = "grey"),
        plot.title =  element_text(color = "black"),
        axis.text = element_text(color = "black"),
        legend.position = "top")
```

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.

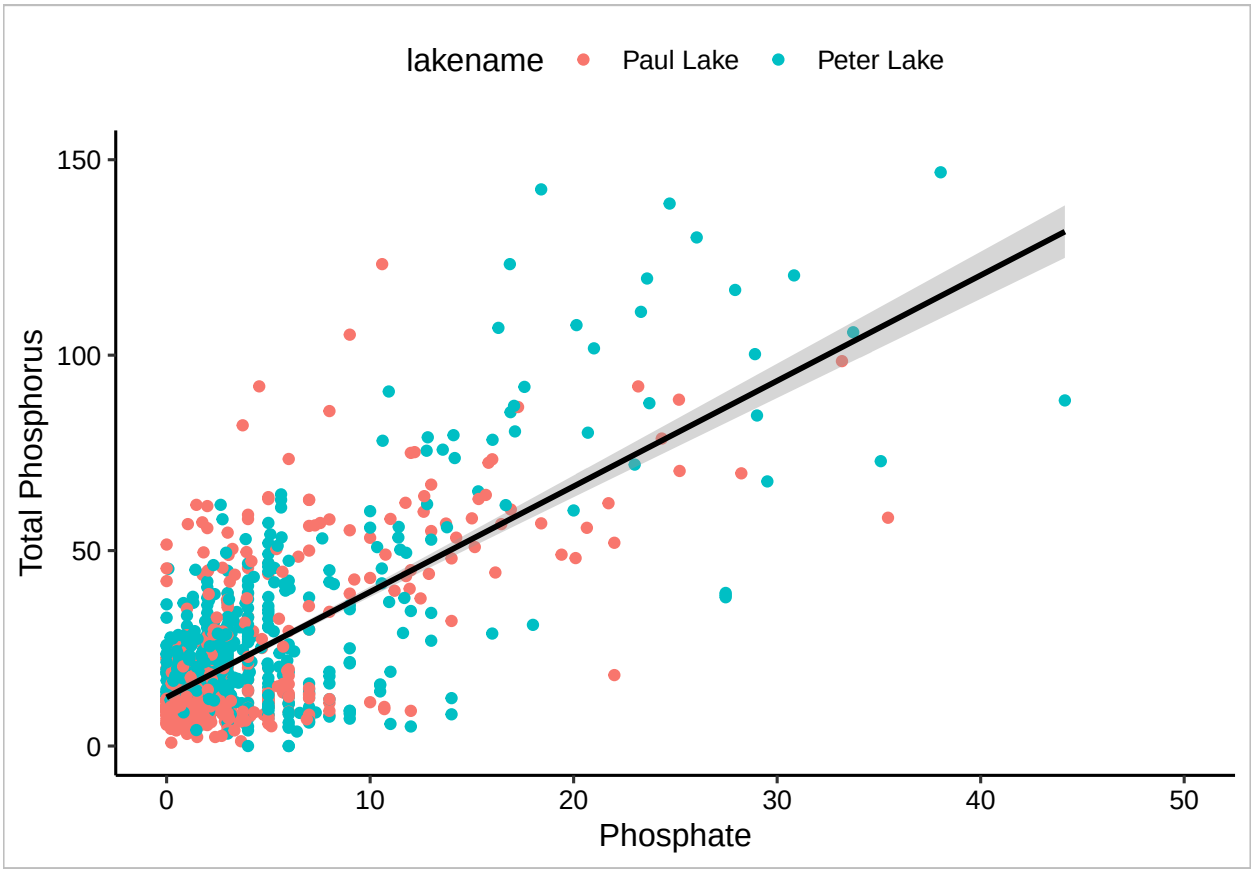
4. [NTL-LTER] Plot total phosphorus (`tp_ug`) by phosphate (`po4`), with separate aesthetics for Peter and Paul lakes. Add a line of best fit and color it black. Adjust your axes to hide extreme values (hint: change the limits using `xlim()` and/or `ylim()`).

```
#4
PeterPaul %>%
  ggplot(aes(x = po4, y = tp_ug, color = lakename)) +
  geom_point() +
  xlim(0, 50) +
  ylim(0, 150) +
  geom_smooth(method = lm, color = "black") +
  labs(x = "Phosphate", y = "Total Phosphorus") +
  mytheme
```

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

```
## Warning: Removed 21948 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 21948 rows containing missing values (`geom_point()`).
```

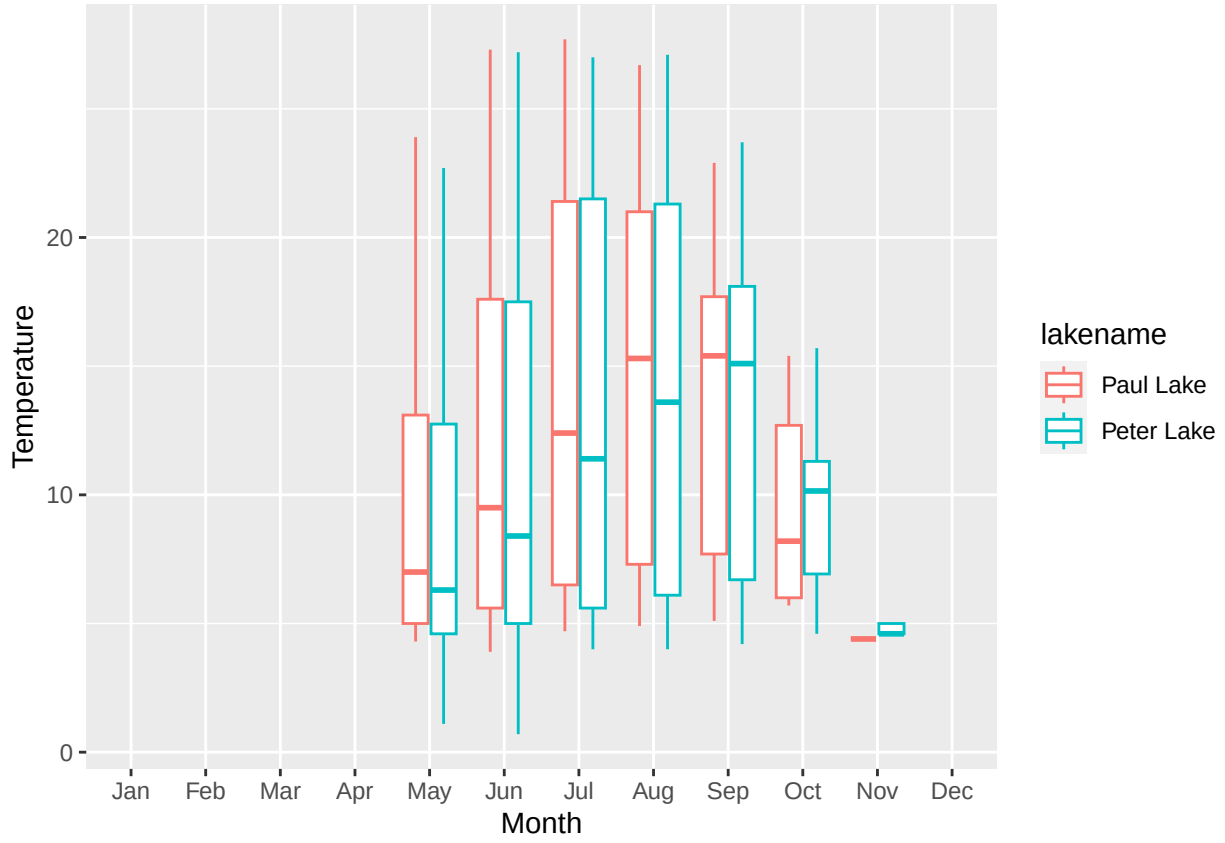


5. [NTL-LTER] Make three separate boxplots of (a) temperature, (b) TP, and (c) TN, with month as the x axis and lake as a color aesthetic. Then, create a cowplot that combines the three graphs. Make sure that only one legend is present and that graph axes are aligned.

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

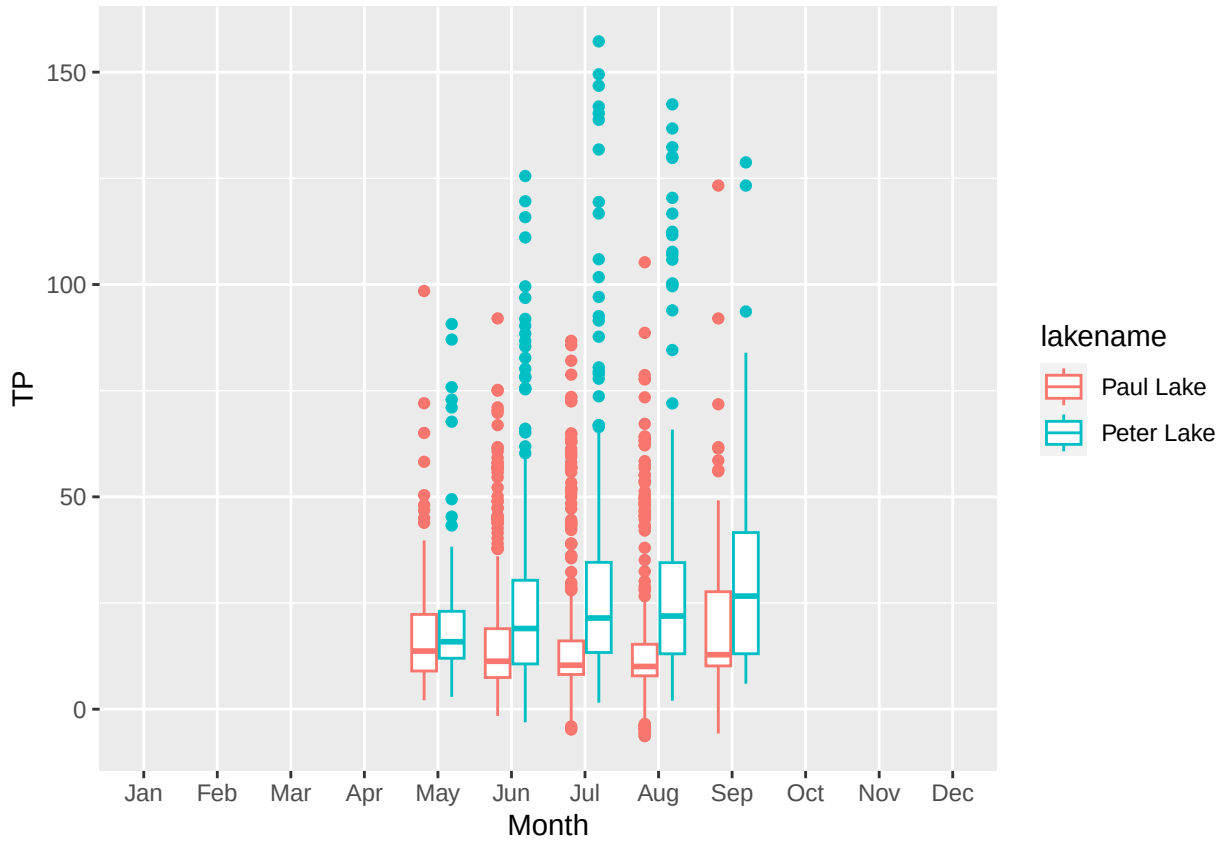
```
#5
temp <- PeterPaul %>%
  ggplot(aes(x = factor(month, levels = 1:12, labels = month.abb),
             y = temperature_C, color = lakename)) +
  geom_boxplot() +
  scale_x_discrete(drop=FALSE) +
  labs(x = "Month", y = "Temperature")
print(temp)
```

```
## Warning: Removed 3566 rows containing non-finite values (`stat_boxplot()`).
```



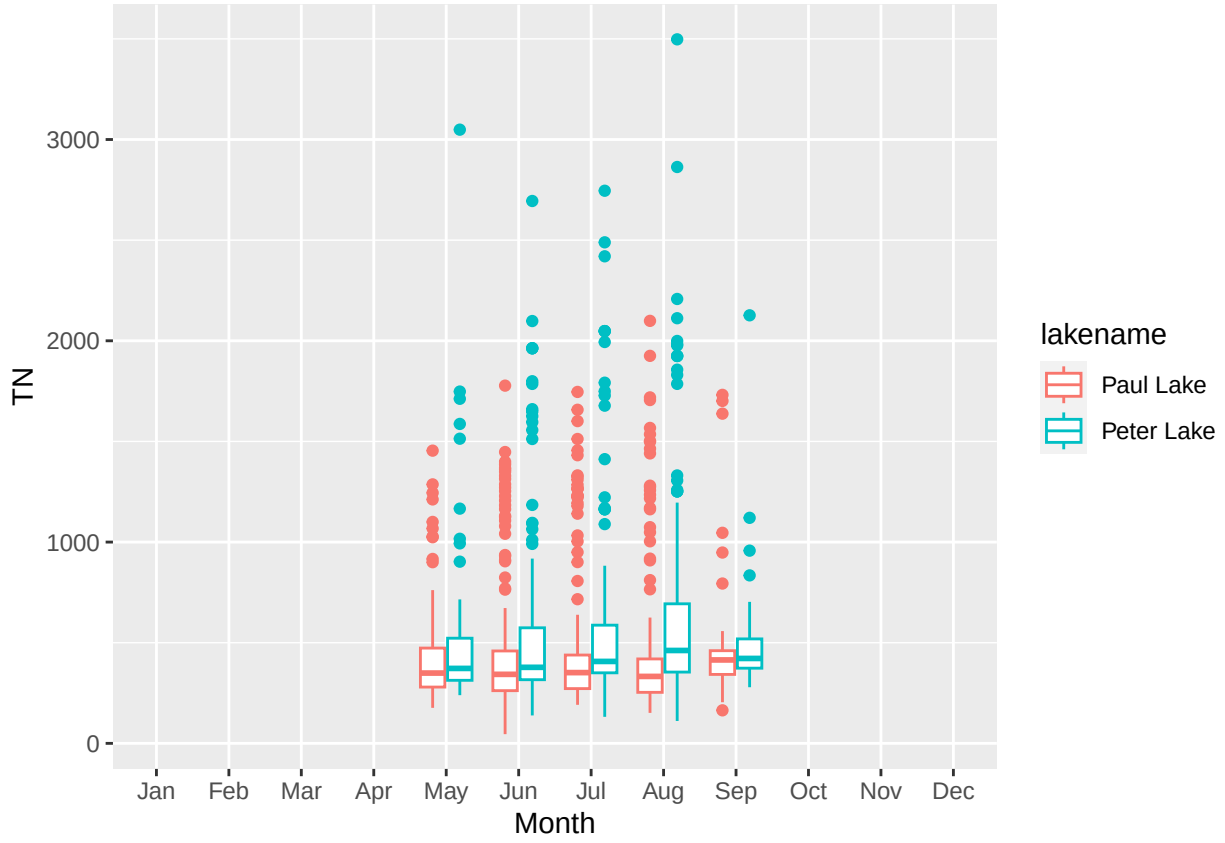

```
tp <- PeterPaul %>%
  ggplot(aes(x = factor(month, levels = 1:12, labels = month.abb),
             y = tp_ug, color = lakename)) +
  geom_boxplot() +
  scale_x_discrete(drop=FALSE) +
  labs(x = "Month", y = "TP")
print(tp)
```

```
## Warning: Removed 20729 rows containing non-finite values (‘stat_boxplot()‘).
```




```
tn <- PeterPaul %>%
  ggplot(aes(x = factor(month, levels = 1:12, labels = month.abb),
             y = tn_ug, color = lakename)) +
  geom_boxplot() +
  scale_x_discrete(drop=FALSE) +
  labs(x = "Month", y = "TN")
print(tn)
```

```
## Warning: Removed 21583 rows containing non-finite values (‘stat_boxplot()‘).
```

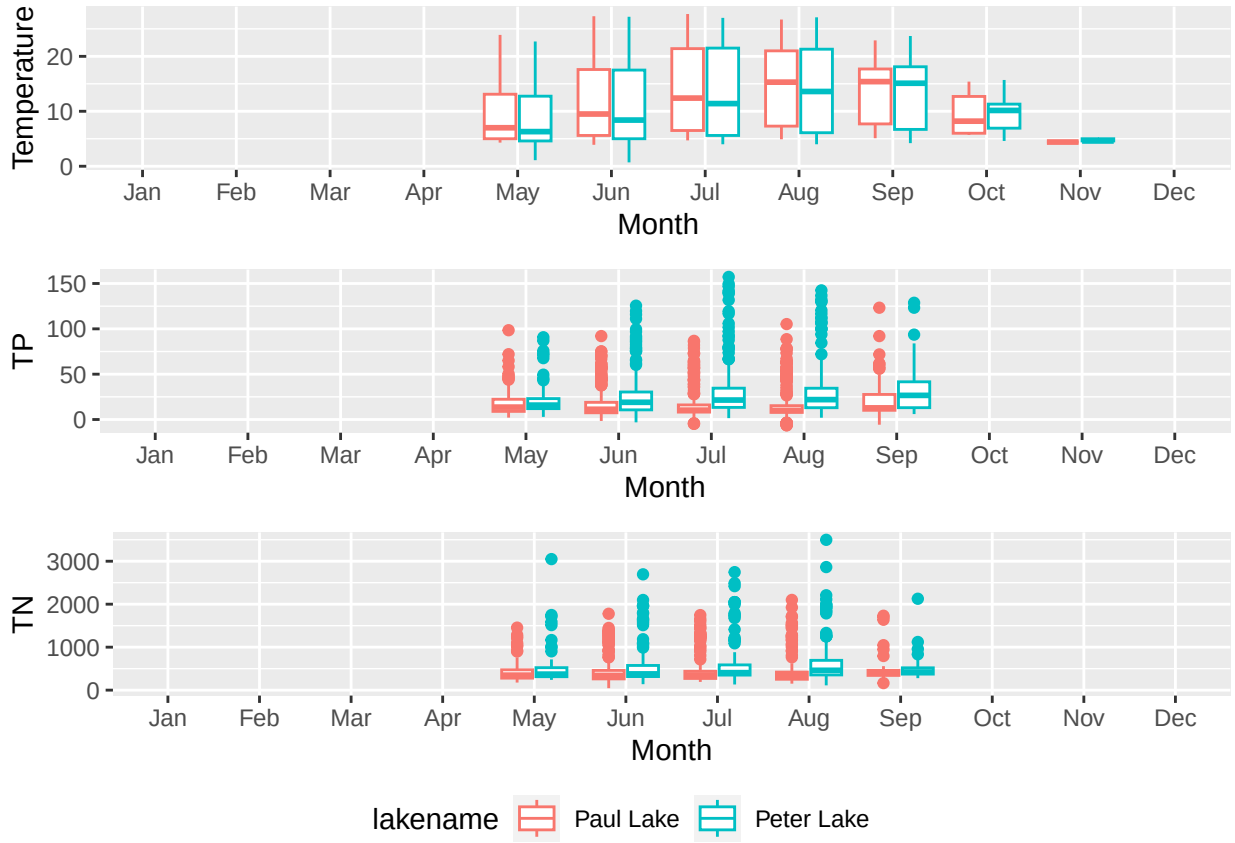



```
ggarrange(temp, tp, tn, nrow = 3, common.legend = T, legend = "bottom")
```

```
## Warning: Removed 3566 rows containing non-finite values (‘stat_boxplot()‘).
## Removed 3566 rows containing non-finite values (‘stat_boxplot()‘).

## Warning: Removed 20729 rows containing non-finite values (‘stat_boxplot()‘).

## Warning: Removed 21583 rows containing non-finite values (‘stat_boxplot()‘).
```



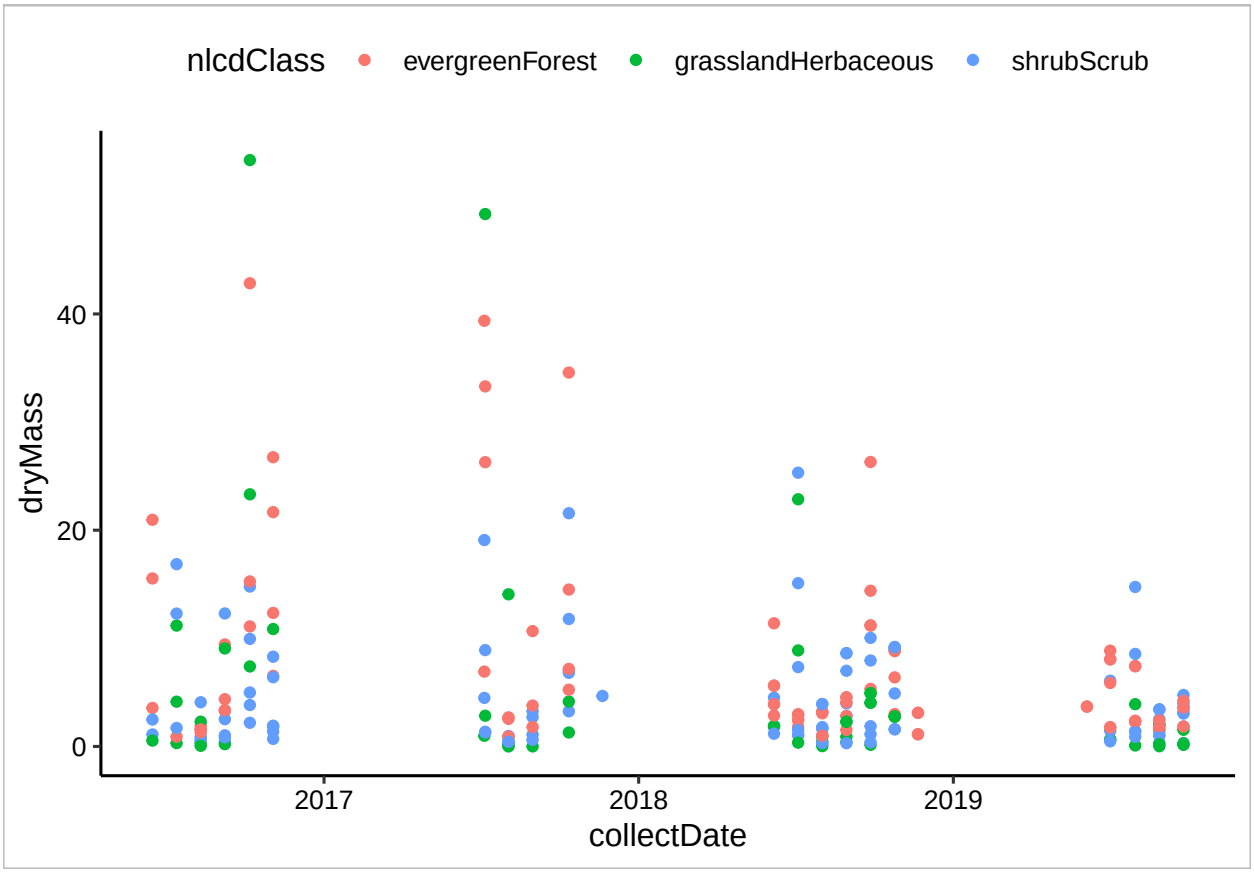
Question: What do you observe about the variables of interest over seasons and between lakes?

> Answer: The temperature for both lakes rises from May to August and September, then it starts to decrease. The Paul Lake generally has a higher temperature than Peter Lake during the spring and summer, then it quickly decreases and becomes lower than the Peter Lake in October. The total phosphorus in Paul Lake is slightly lower from June to August and is always lower than that in the Peter Lake. Peter Lake has a relatively higher TP in the summer. The total nitrogen's seasonal change is weak. But Paul Lake is always lower than the Peter Lake.

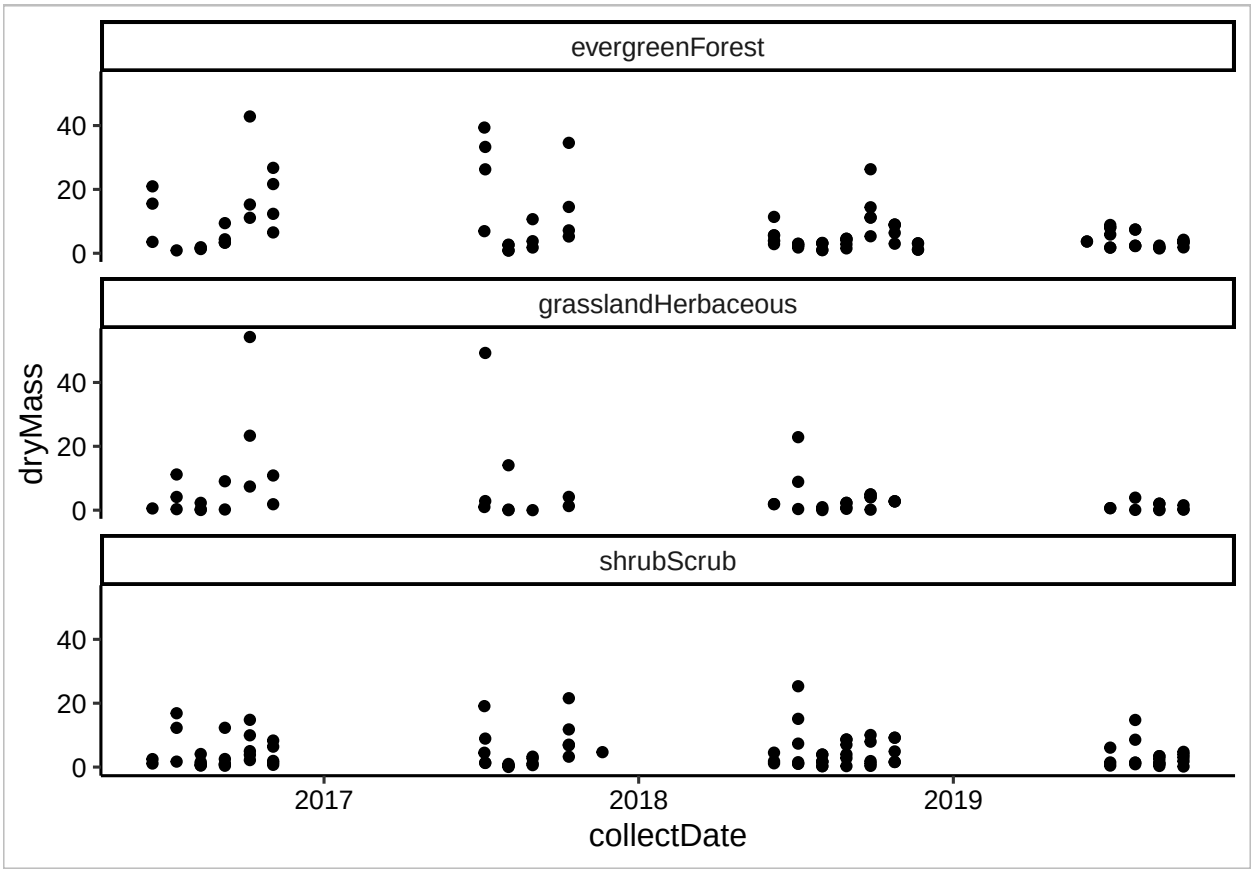
6. [Niwot Ridge] Plot a subset of the litter dataset by displaying only the "Needles" functional group. Plot the dry mass of needle litter by date and separate by NLCD class with a color aesthetic. (no need to adjust the name of each land use)

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

```
#6
needles <- Niwot %>%
  filter(functionalGroup == "Needles") %>%
  ggplot(aes(x = collectDate, y = dryMass, color = nlcdClass)) +
  geom_point() +
  mytheme
print(needles)
```




```
#7
needles_nlcd <- Niwot %>%
  filter(functionalGroup == "Needles") %>%
  ggplot(aes(x = collectDate, y = dryMass)) +
  geom_point() +
  facet_wrap(vars(nlcdClass), nrow = 3) +
  mytheme
print(needles_nlcd)
```



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

Answer:If we want to compare the three NLCD classes, I feel graph 6 separated by color is more effective. Because most of the needles dry mass has pretty similar distributions in the three NLCD classes and it's hard to compare them when they are in three separate graphs. If we want to know the dry mass and distribution for each NLCD class, then graph 7 is better. It's clear to see the pattern over time and for each class when they are separated.