Assignment 5: Data Visualization

Devin Domeyer

OVERVIEW

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

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_A05_DataVisualization.Rmd") prior to submission.

The completed exercise is due on Monday, February 14 at 7:00 pm.

Set up your session

1. Set up your session. Verify your working directory and load the tidyverse and 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_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] "/Users/devindomeyer/Desktop/Duke/Data Analytics/Environmental_Data_Analytics_2022"
```

```
#install.packages("tidyverse")
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#install.packages("cowplot")
library(cowplot)
library(ggplot2)
#install.packages("cowplot")
library(cowplot)

lake <- read.csv("./Data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv", stringsA
litter <- read.csv("./Data/Processed/NEON_NIWO_Litter_mass_trap_Processed.csv", stringsAsFactors = TRUE
#2
class(lake$sampledate)
```

```
## [1] "factor"
```

```
lake$sampledate <- as.Date(lake$sampledate, format = "%Y-%m-%d")
class(lake$sampledate)
```

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

```
lake$month <- as.factor(lake$month)

class(litter$collectDate)
```

```
## [1] "factor"
```

```
litter$collectDate <- as.Date(litter$collectDate, format = "%Y-%m-%d")
class(litter$collectDate)
```

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

Define your theme

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

```
#3
mytheme <- theme_light(base_size = 14) +
  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.

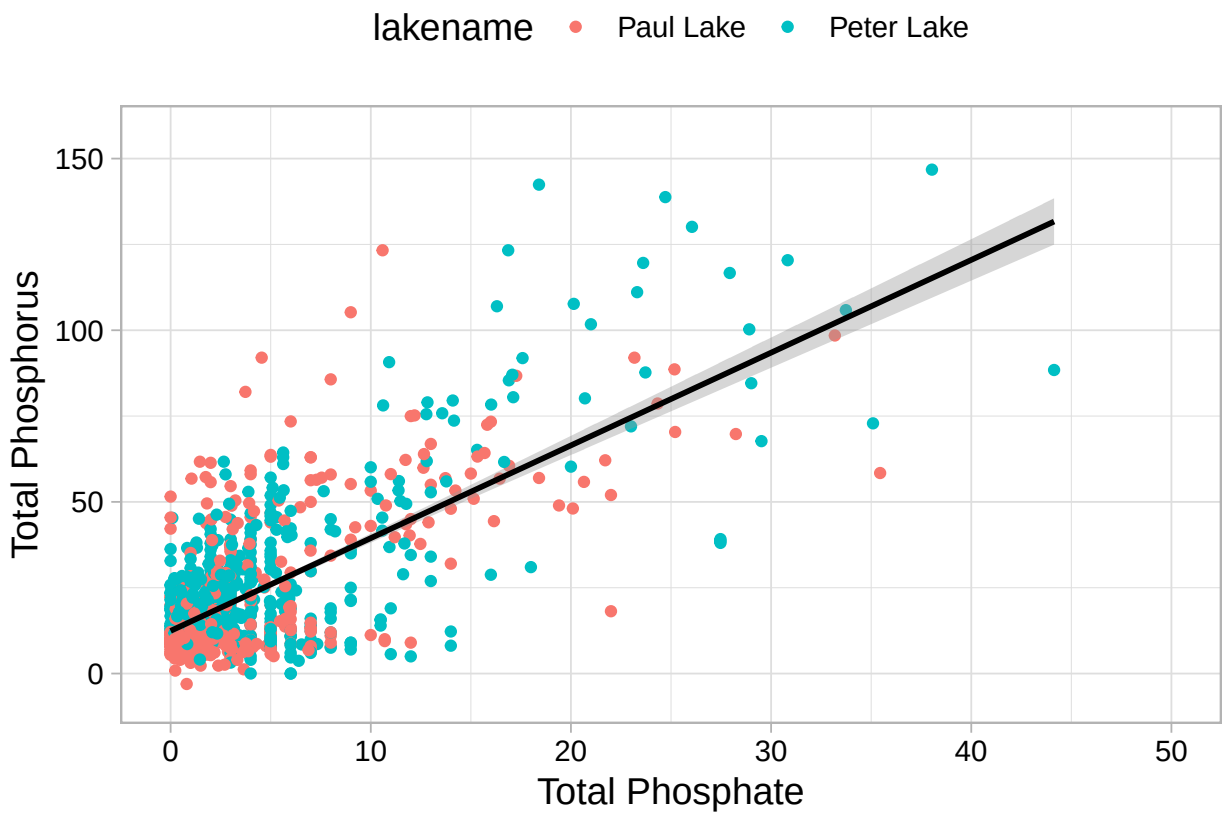
4. [NTL-LTER] Plot total phosphorus (`tp_ug`) by phosphate (`po4`), with separate aesthetics for Peter and Paul lakes. Add a line of best fit and color it black. Adjust your axes to hide extreme values (hint: change the limits using `xlim()` and `ylim()`).

```
#4
lakeplot1 <- ggplot(lake, aes(x=po4, y=tp_ug)) +
  geom_point(aes(color=lakename))+
  geom_smooth(method=lm, color="black")+
  xlim(0, 50)+
  ylab("Total Phosphorus")+
  xlab("Total Phosphate")
print(lakeplot1)
```

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

## Warning: Removed 21947 rows containing non-finite values (stat_smooth).

## Warning: Removed 21947 rows containing missing values (geom_point).
```

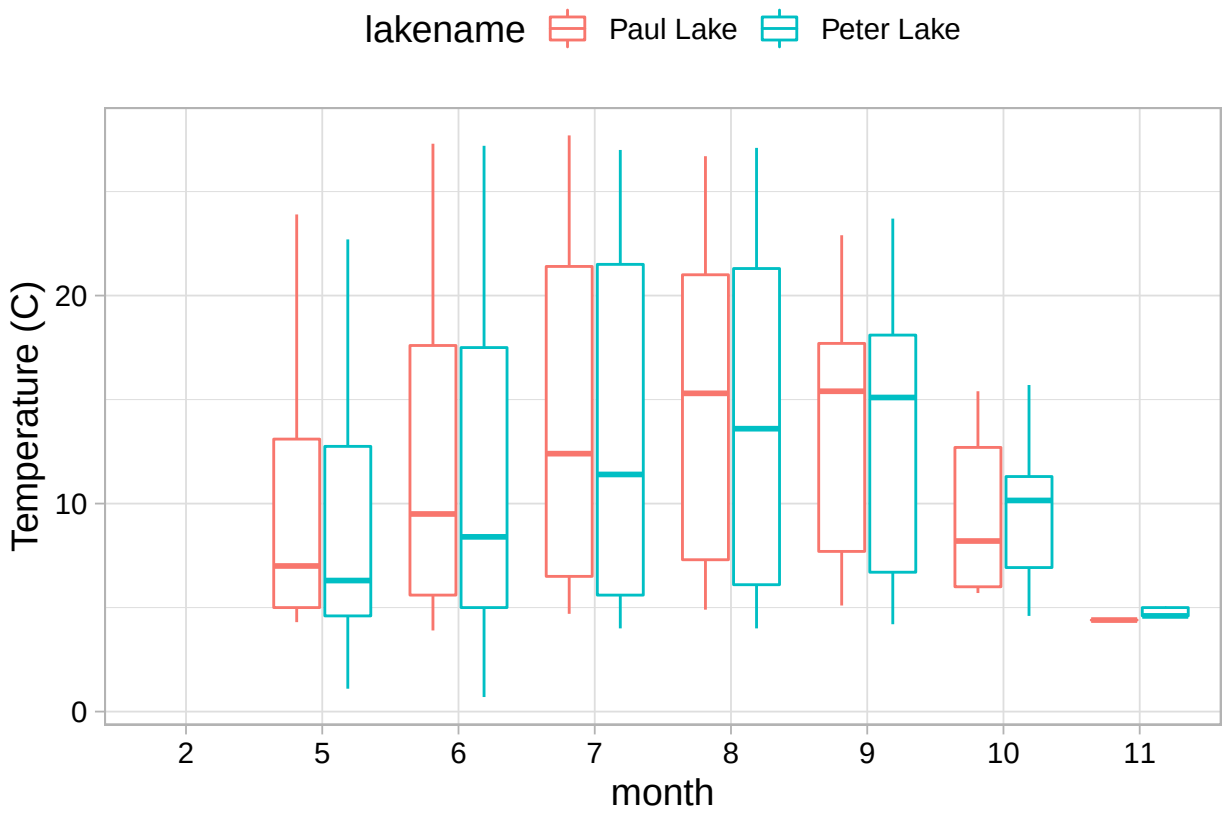


5. [NTL-LTER] Make three separate boxplots of (a) temperature, (b) TP, and (c) TN, with month as the x axis and lake as a color aesthetic. Then, create a cowplot that combines the three graphs. Make sure that only one legend is present and that graph axes are aligned.

```
#5
lakeplot2 <- ggplot(lake, aes(x=month, y=temperature_C))+
  geom_boxplot(aes(color=lakename))+
```

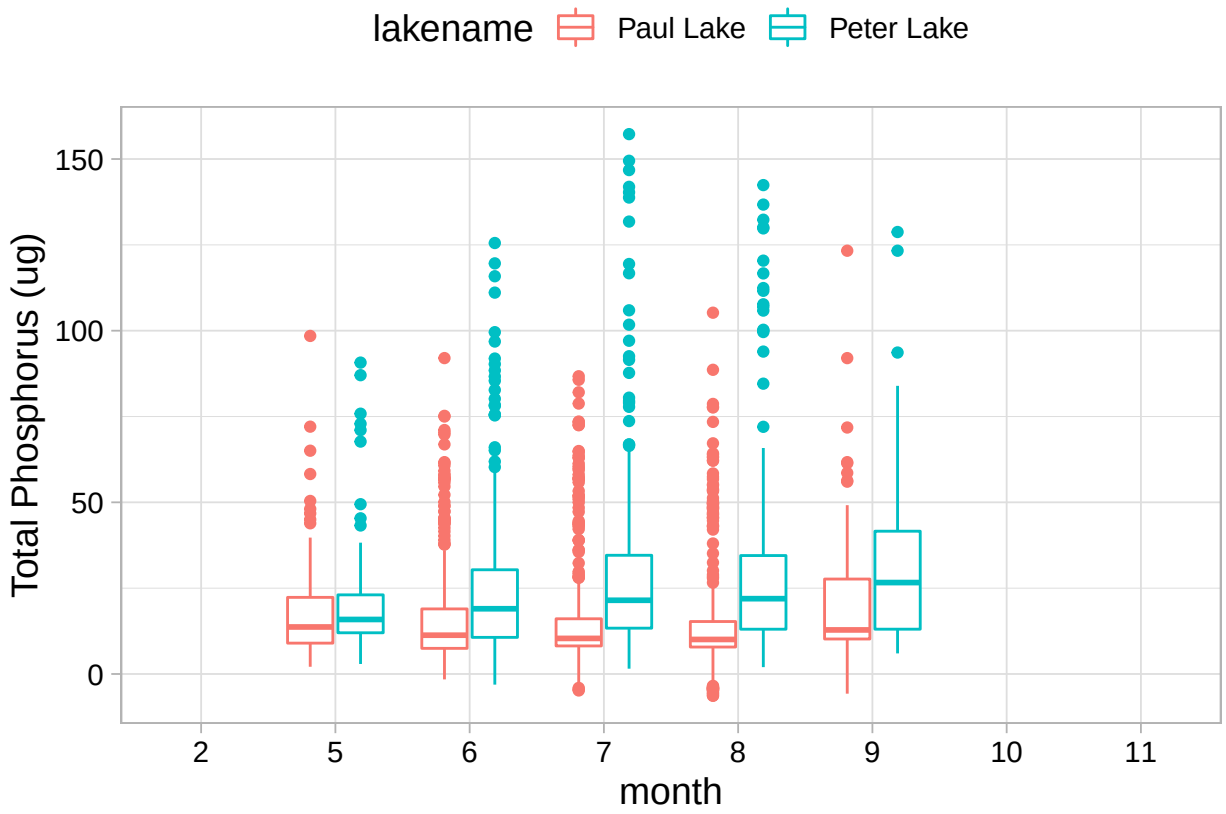
```
  ylab("Temperature (C)")
print(lakeplot2)
```

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



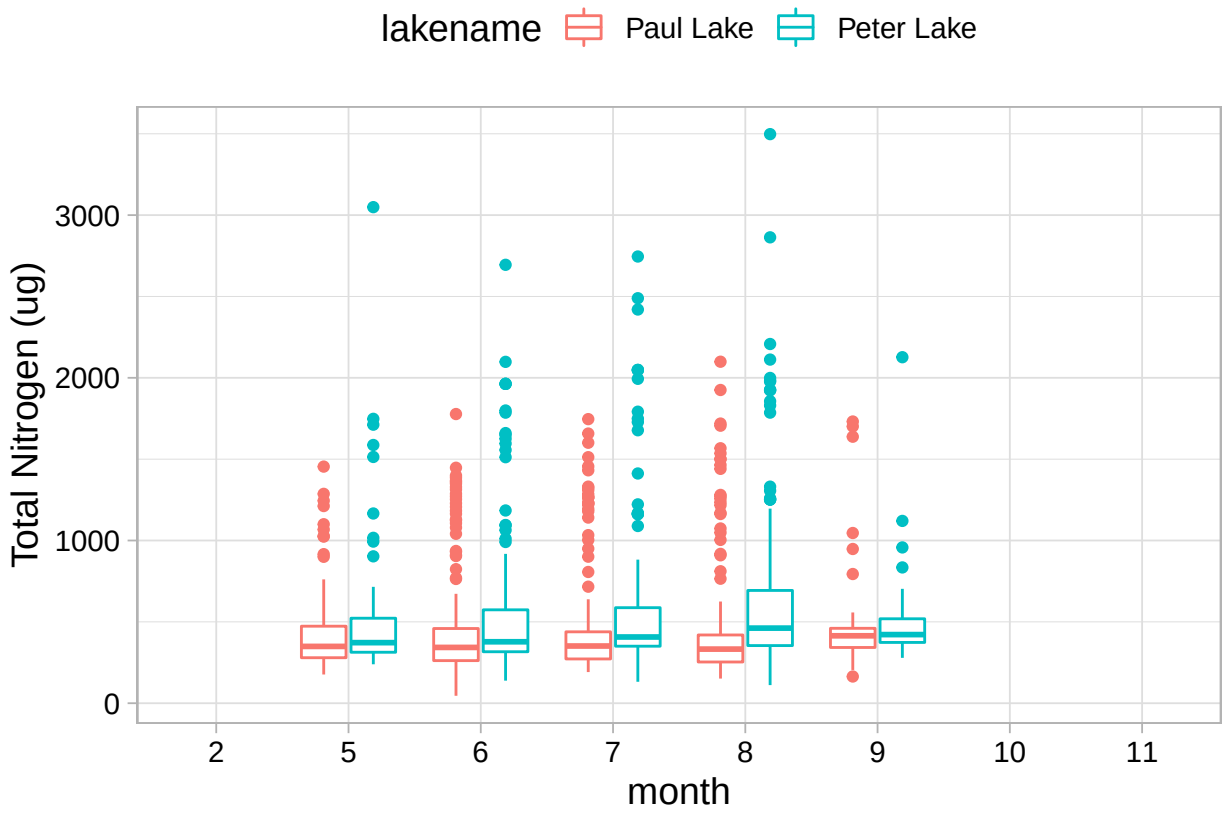

```
lakeplot3 <- ggplot(lake, aes(x=month, y=tp_ug))+
  geom_boxplot(aes(color=lakename))+
  ylab("Total Phosphorus (ug)")
print(lakeplot3)
```

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




```
lakeplot4 <- ggplot(lake, aes(x=month, y=tn_ug))+
  geom_boxplot(aes(color=lakename))+
  ylab("Total Nitrogen (ug)")
print(lakeplot4)
```

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




```
lakegrid <- plot_grid(lakeplot2 + theme(legend.position="none"), lakeplot3 + theme(legend.position="none
```

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

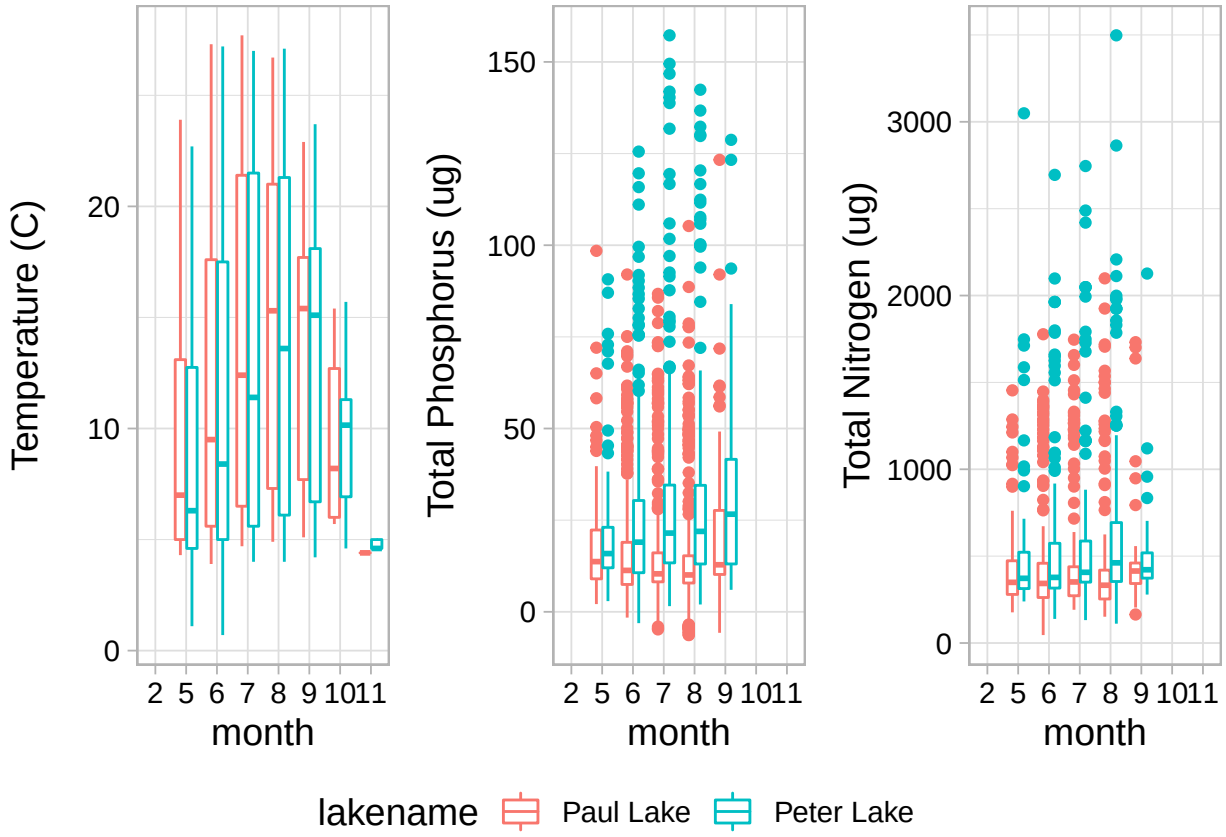
```
## Warning: Removed 20729 rows containing non-finite values (stat_boxplot).
```

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

```
legend <- get_legend(lakeplot2 +
    guides(color = guide_legend(nrow = 1)) +
    theme(legend.position = "bottom"))
```

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

```
plot_grid(lakegrid, legend, ncol=1, rel_heights = c(1, .1))
```



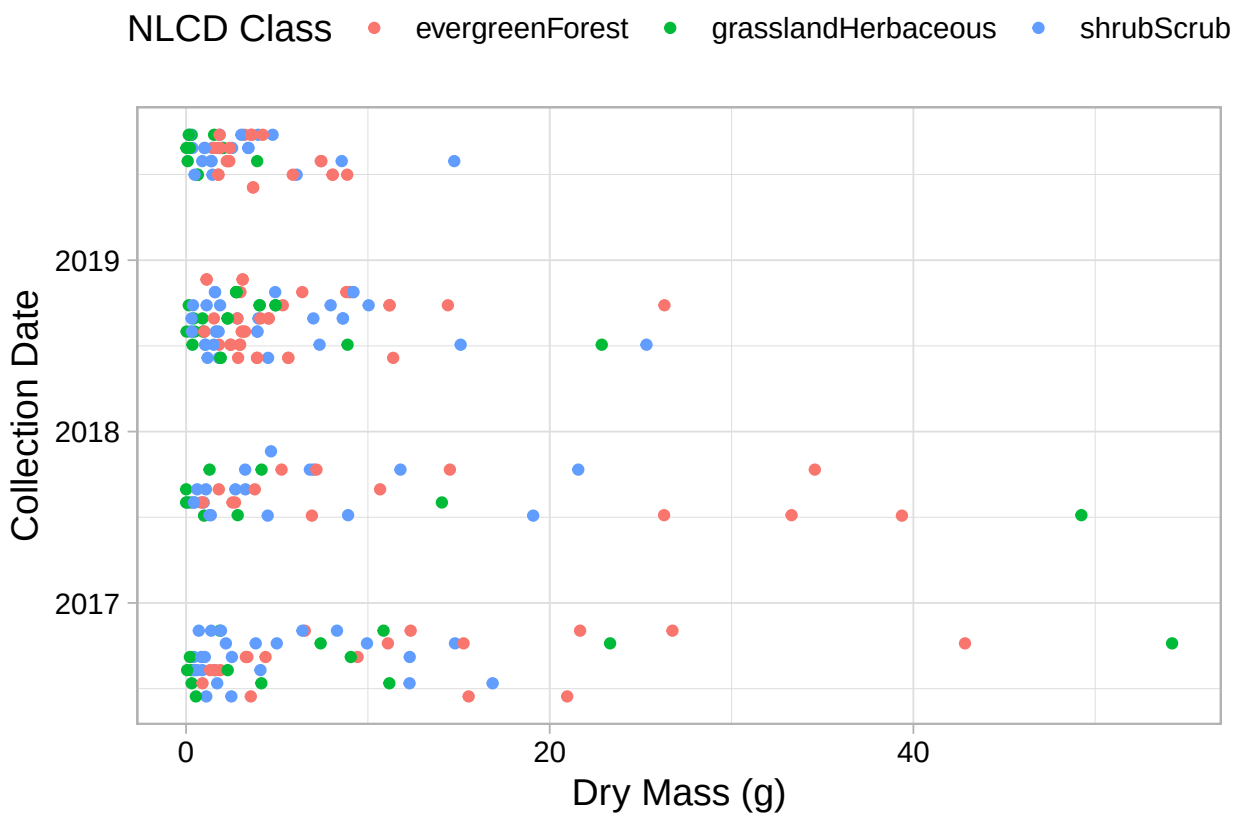
Question: What do you observe about the variables of interest over seasons and between lakes?

> Answer: Across months,Peter Lakehad higher average total nitrogen and total phosphorus and a wider spread of concentrations for these nutrients. Paul lake had higher average temperatures across months with the exception of October and November, in which Peter Lake had higher averages. Data for total phosphorus and total nitrogen were not collected in October and November.

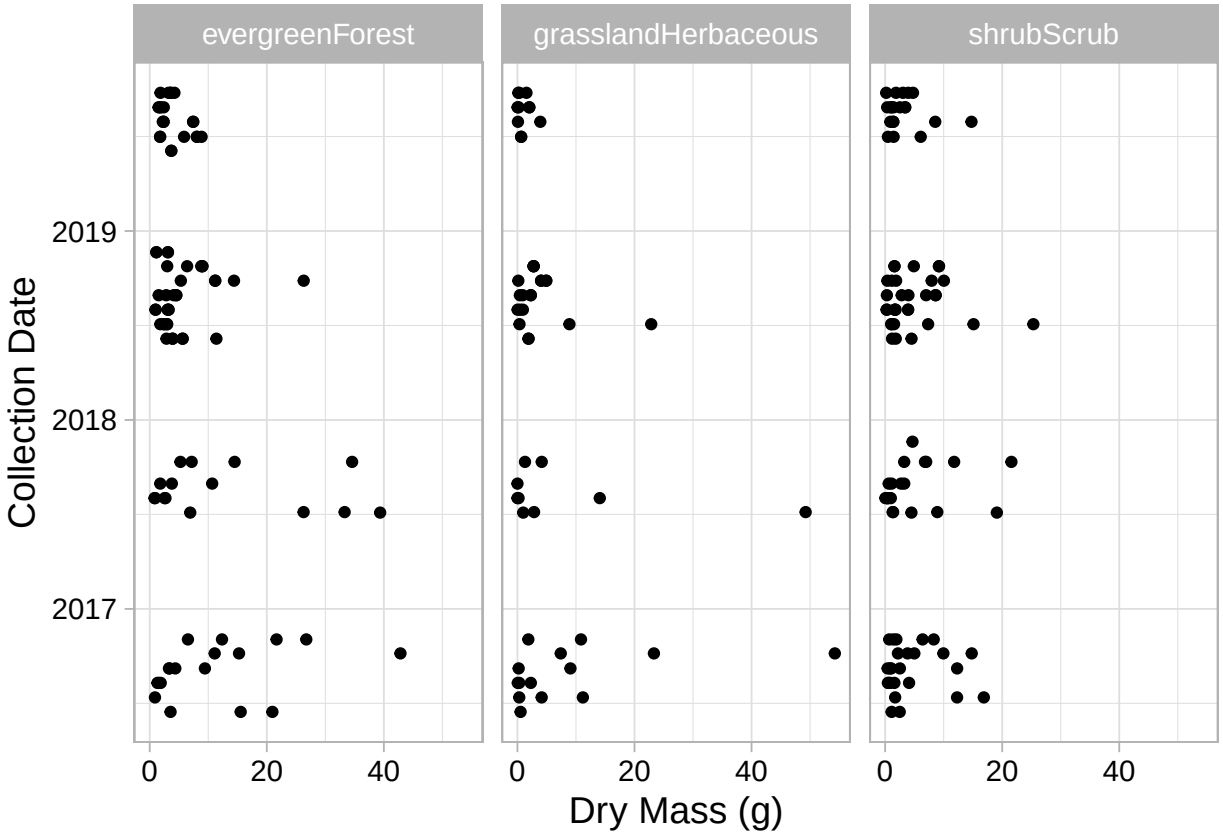
6. [Niwot Ridge] Plot a subset of the litter dataset by displaying only the "Needles" functional group. Plot the dry mass of needle litter by date and separate by NLCD class with a color aesthetic. (no need to adjust the name of each land use)

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

```
#6
litterplot1 <- ggplot(subset(litter, functionalGroup == "Needles"), aes(x=dryMass, y=collectDate))+
  geom_point(aes(color=nlcdClass))+
  ylab("Collection Date")+
  xlab("Dry Mass (g)")+
  labs(color="NLCD Class")
print(litterplot1)
```




```
#7
litterplot2 <- ggplot(subset(litter, functionalGroup == "Needles"), aes(x=dryMass, y=collectDate))+
  geom_point()+
  facet_wrap(vars(nlcdClass), ncol=3)+
  ylab("Collection Date")+
  xlab("Dry Mass (g)")
print(litterplot2)
```



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

Answer:I think that 7 is more effective because it's easy to differentiate between the different NLCD classes. In contrast, with plot 6 you have to pick out each NCLP class by point color. Plot 7 is better for making inferences about differences in dry mass between classes.