Assignment 5: Data Visualization

Christina Li

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 Tuesday, February 23 at 11:59 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 (both the tidy [`NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv`] and the gathered [`NTL-LTER_Lake_Nutrients_PeterPaulGathered_Processed.csv`] versions) and the processed data file for the Niwot Ridge litter dataset.

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

```
# 1
getwd()
```

```
## [1] "C:/Users/li_ch/Desktop/DKU/Year 2/Term 2/Environmental Data Analytics/GIT Hub/Environmental_Data
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.5     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

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

```
library(cowplot)

# 2
Chemistry <- read.csv("./Data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv",
    stringsAsFactors = TRUE)
```

```
Gathered <- read.csv("./Data/Processed/NTL-LTER_Lake_Nutrients_PeterPaulGathered_Processed.csv",
    stringsAsFactors = TRUE)
Litter <- read.csv("./Data/Processed/NEON_NIWO_Litter_mass_trap_Processed.csv",
    stringsAsFactors = TRUE)
```

Define your theme

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

```
Theme_1 <- theme_bw(base_size = 11) + theme(plot.title = element_text(face = "bold",
    size = 12, hjust = 0.5), axis.text.x = element_text(vjust = 1, hjust = 0),
    legend.margin = margin(1), legend.position = "right", legend.text = element_text(size = 11),
    legend.title = element_text(size = 11))

theme_set(Theme_1)
```

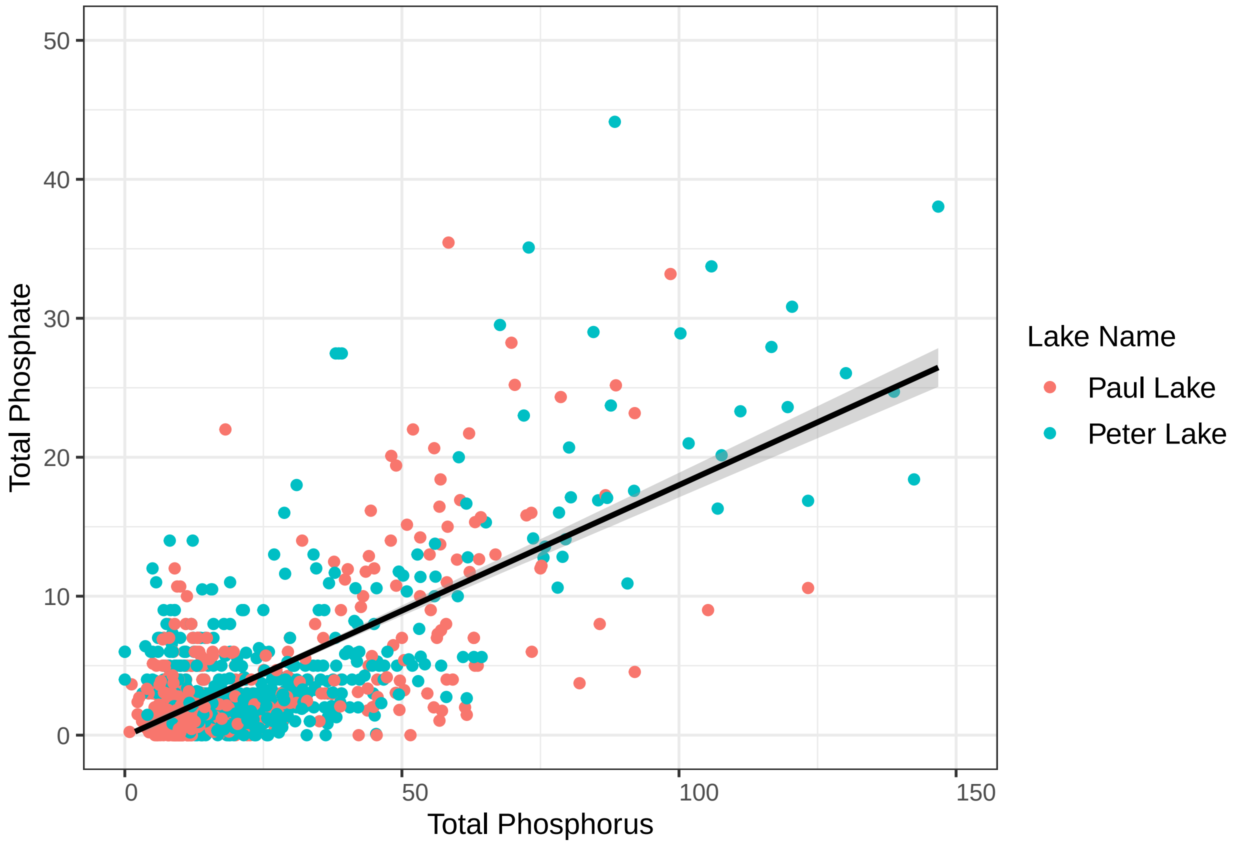
Create graphs

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.

```
TP_by_Phosphate <- ggplot(Chemistry, aes(x = tp_ug, y = po4, color = lakename)) +
    geom_point() + xlim(0, 150) + ylim(0, 50) + labs(x = "Total Phosphorus",
    y = "Total Phosphate", color = "Lake Name") + geom_smooth(method = lm,
    color = "black")
print(TP_by_Phosphate)
```

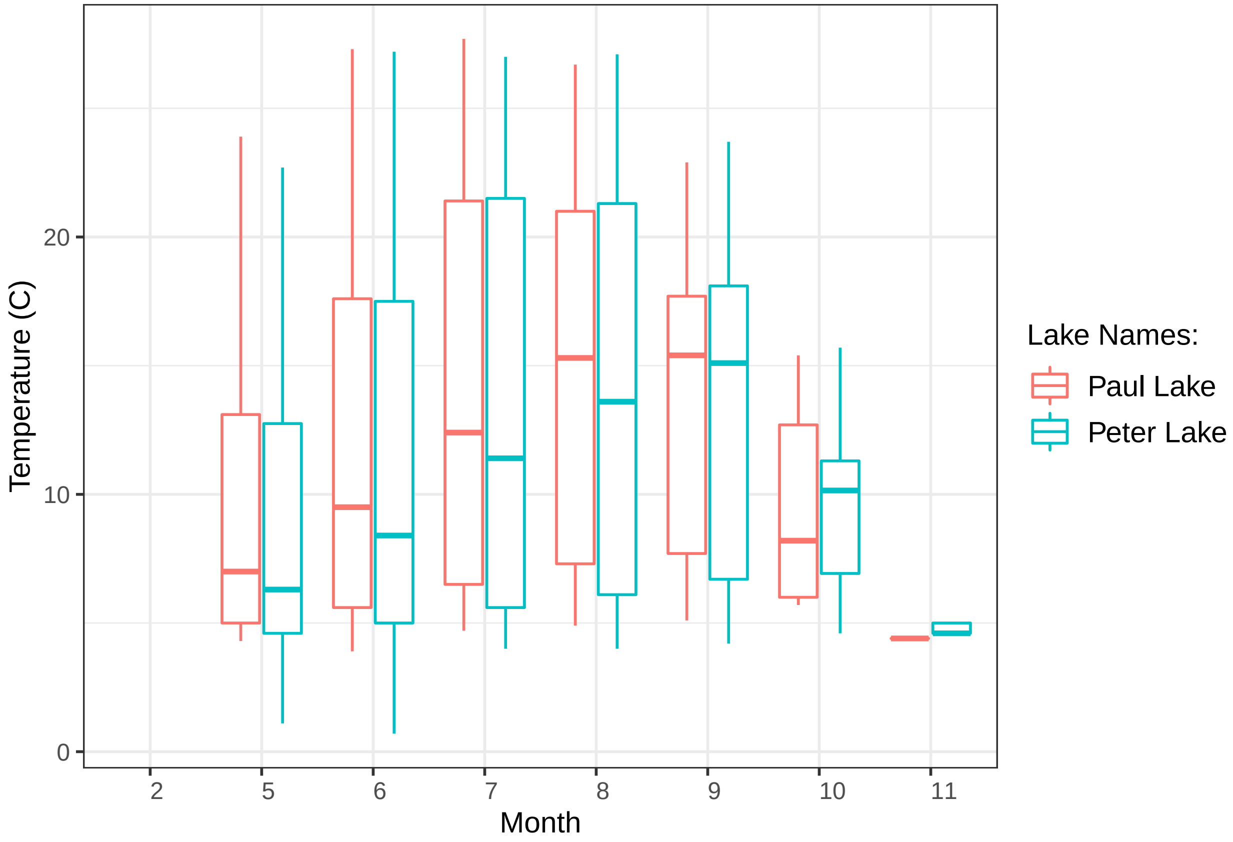
```
## `geom_smooth()` using formula 'y ~ x'

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

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

## Warning: Removed 1 rows containing missing values (geom_smooth).
```



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 combinesclass(Chemistry$month) the three graphs. Make sure that only one legend is present and that graph axes are aligned.

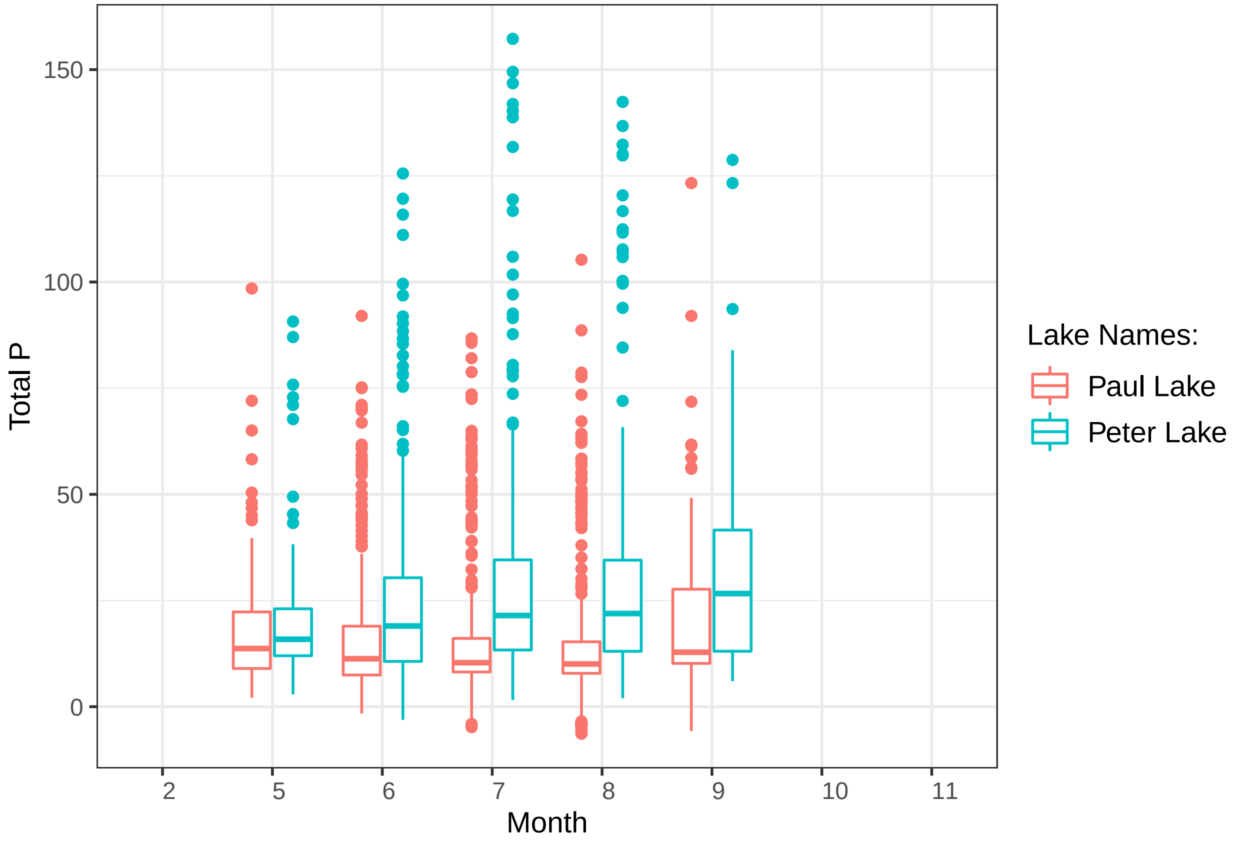
```
Chemistry$month <- as.factor(Chemistry$month)

Temperature <- ggplot(Chemistry) + geom_boxplot(aes(x = month, y = temperature_C,
    color = lakename)) + labs(x = "Month", y = "Temperature (C)", color = "Lake Names: ")
print(Temperature)
```

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



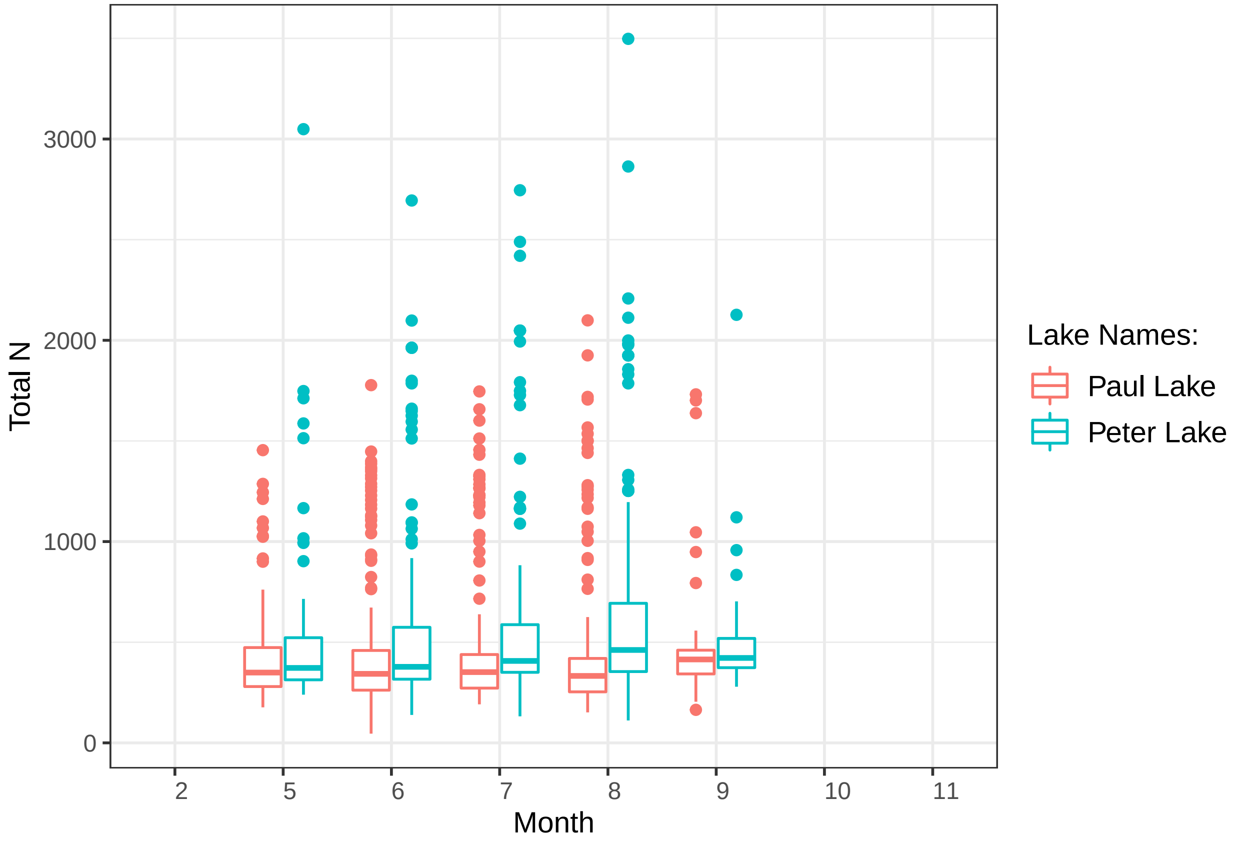
```
TP <- ggplot(Chemistry) + geom_boxplot(aes(x = month, y = tp_ug, color = lakename)) +
    labs(x = "Month", y = "Total P", color = "Lake Names: ")
print(TP)
```

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



```
TN <- ggplot(Chemistry) + geom_boxplot(aes(x = month, y = tn_ug, color = lakename)) +
    labs(x = "Month", y = "Total N", color = "Lake Names: ")
print(TN)
```

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



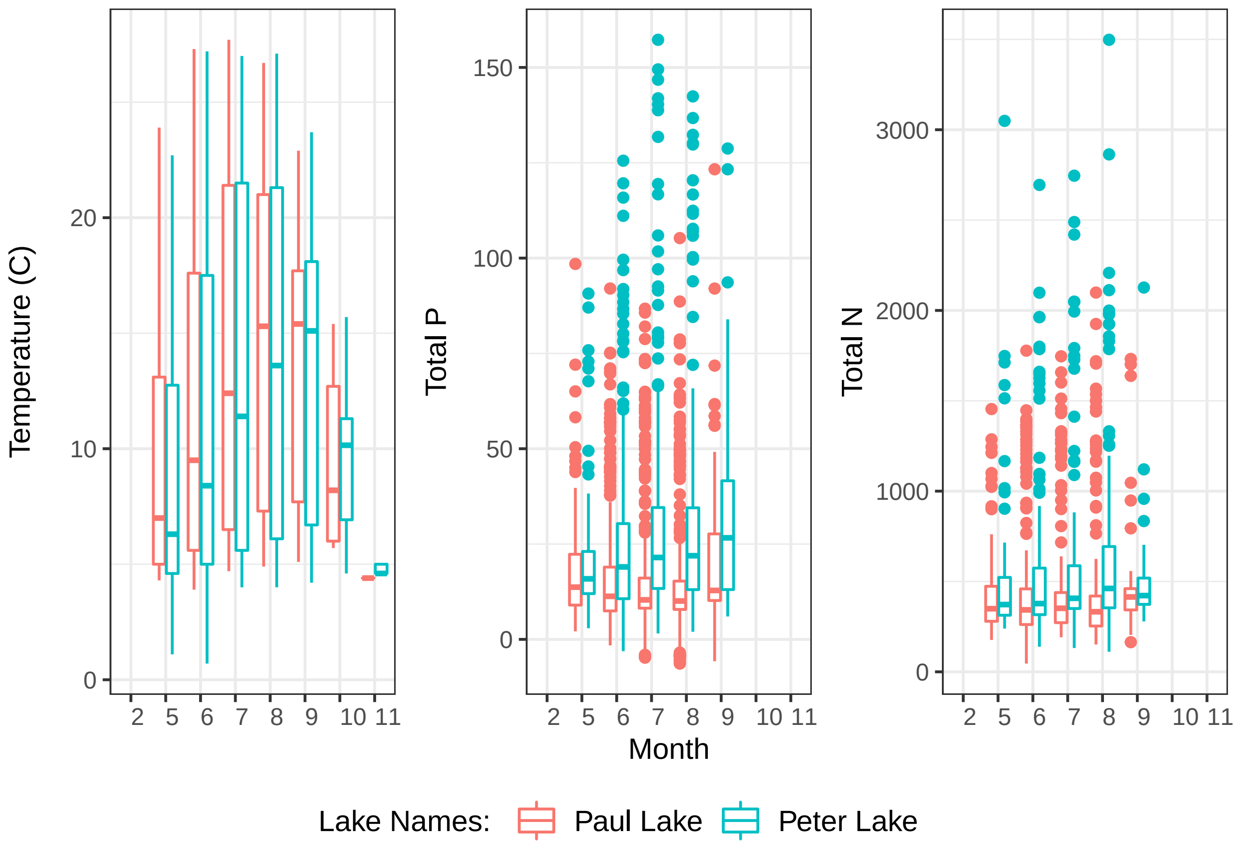
```
Three_in_one <- plot_grid(Temperature + theme(legend.position = "none") +
    xlab(NULL), TP + theme(legend.position = "none"), TN + theme(legend.position = "none") +
    xlab(NULL), nrow = 1, align = "vh")
```

```
## 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(TN + theme(legend.position = "bottom"))
```

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

```
final <- plot_grid(Three_in_one, legend, ncol = 1, rel_heights = c(10,
    1))
print(final)
```

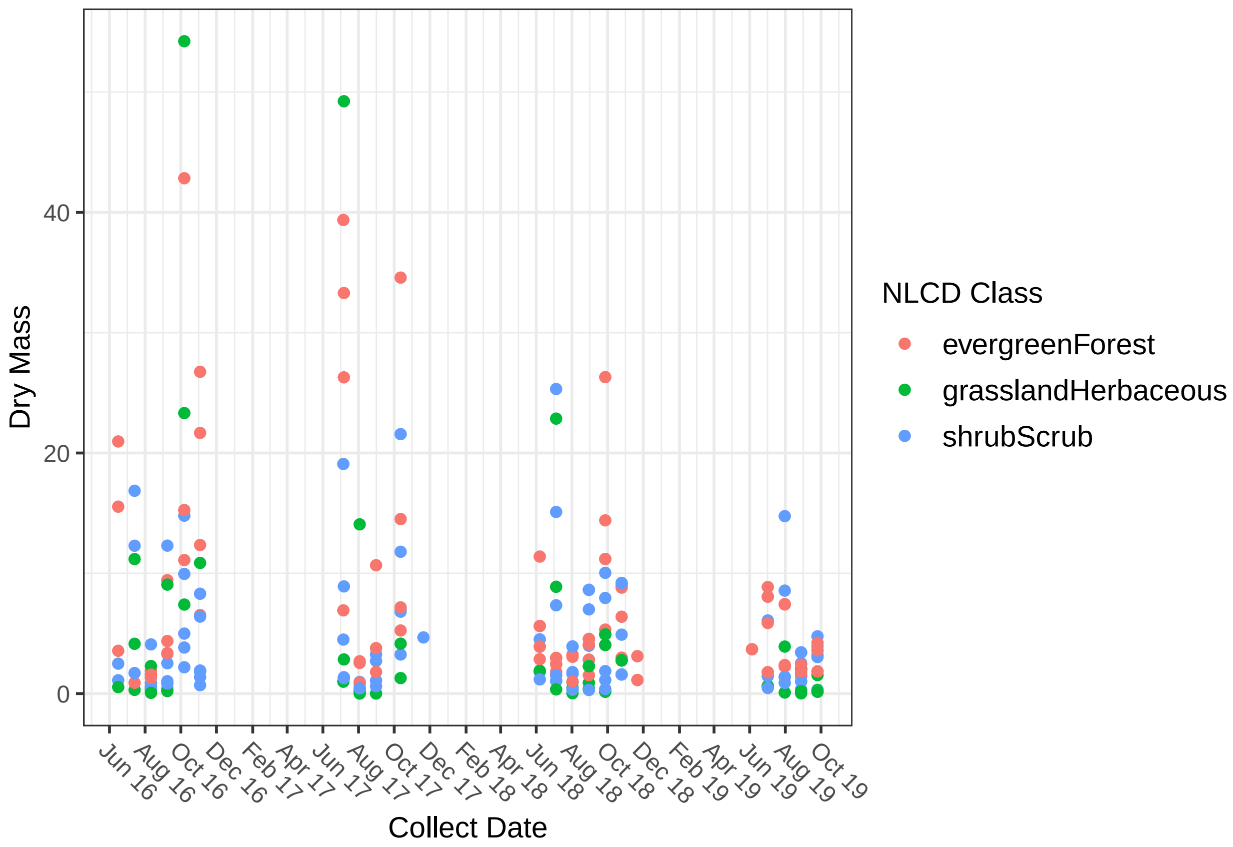


Question: What do you observe about the variables of interest over seasons and between lakes?

> Answer: Total Nitrogen does not vary much over the seasons, while temperature and total phosphorus increase in the summar and decrease in the winter. For both Paul Lake and Peter Lake, the total phosphorus and total nigtogen are higher. However, the temperature of the two lakes are similar.

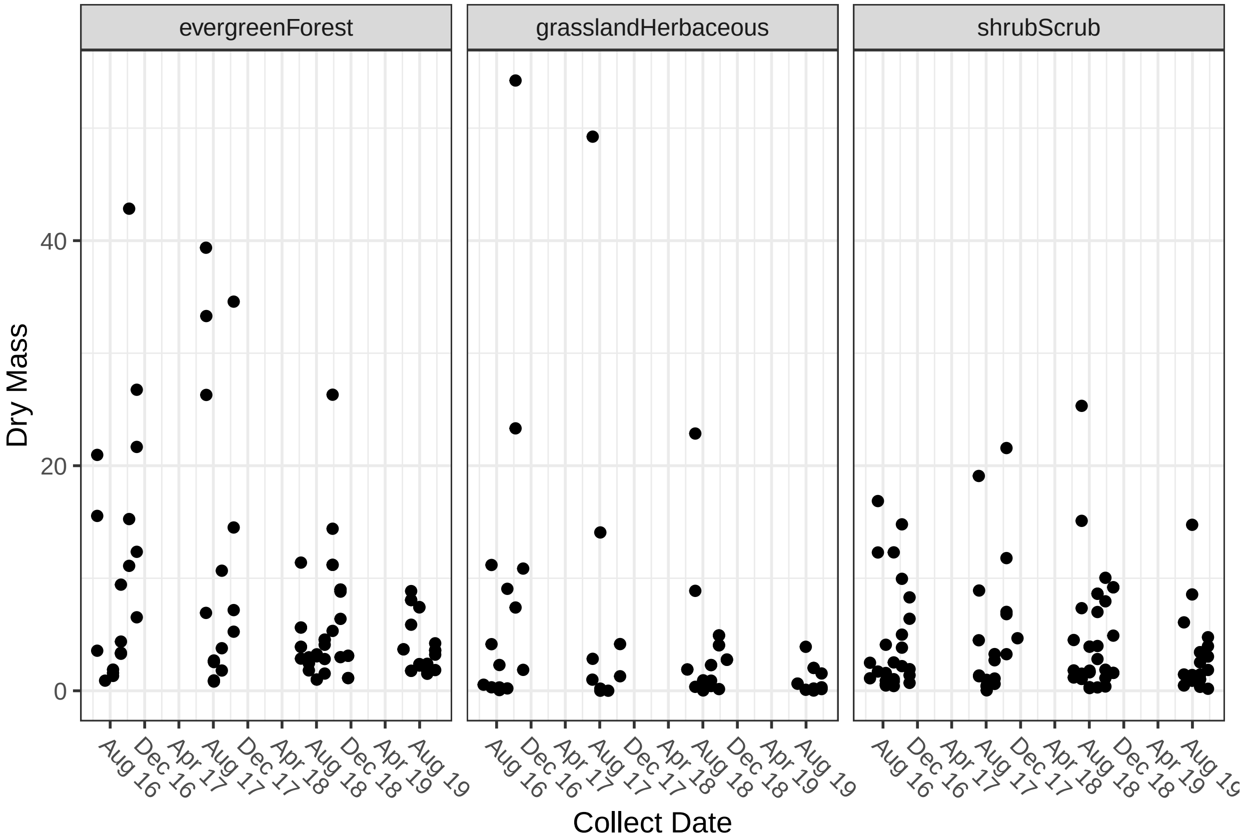
6. [Niwot Ridge] Plot a subset of the litter dataset by displaying only the “Needles” functional group. Plot the dry mass of needle litter by date and separate by NLCD class with a color aesthetic. (no need to adjust the name of each land use)

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

ggplot(subset(Litter, functionalGroup == "Needles")) + geom_point(aes(x = collectDate,
    y = dryMass, color = nlcdClass)) + scale_x_date(date_breaks = "2 months",
    date_labels = "%b %y") + labs(x = "Collect Date", y = "Dry Mass", color = "NLCD Class") +
    theme(axis.text.x = element_text(angle = -45))
```



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

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

ggplot(subset(Litter, functionalGroup == "Needles")) + geom_point(aes(x = collectDate,
    y = dryMass)) + scale_x_date(date_breaks = "4 months", date_labels = "%b %y") +
    facet_wrap(vars(nlcdClass), ncol = 3) + labs(x = "Collect Date", y = "Dry Mass") +
    theme(axis.text.x = element_text(angle = -45))
```



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

> Answer: I chose scatter plot for question 6 and 7 because we are comparing the litter amount collected over time. I think Q6 is more effective if we want to compare the litter dry mass within a year among three NLCD class. However, if we want to compare the variation between each year of the same NLCD class, 7 is better.