Assignment 5: Data Visualization

Ariel Lam

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_PeterPau` version) and the processed data file for the Niwot Ridge litter dataset (use the [`NEON_NIWO_Litter_mass_trap_Processed` version).
2. Make sure R is reading dates as date format; if not change the format to date.

```
# 1
getwd()
```

```
## [1] "/Users/ariellam/Desktop/EDA-Fall2022/Assignments"
```

```
setwd("/Users/ariellam/Desktop/EDA-Fall2022")
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
```

```
library(cowplot)
```

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

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

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

Define your theme

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

```
# 3
assignment_theme <- theme_light(base_size = 12) +
    theme(axis.text = element_text(color = "black"))

theme_set(assignment_theme)
```

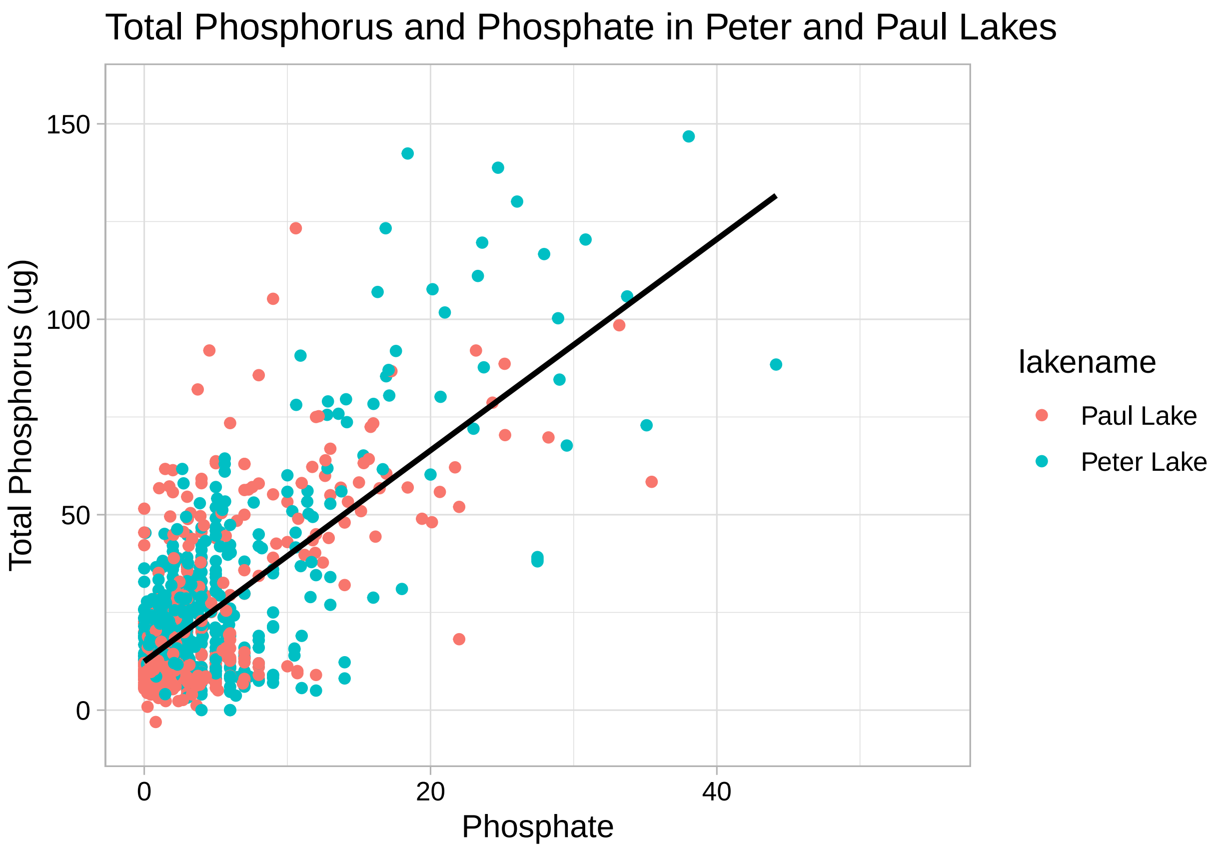
Create graphs

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(PeterPaul, aes(x = po4,
    y = tp_ug)) + geom_point(aes(color = lakename)) +
    xlim(c(0, 55)) +
    geom_smooth(method = lm,
        color = "black",
        se = FALSE) +
    ylab("Total Phosphorus (ug)") +
    xlab("Phosphate") +
    ggtitle("Total Phosphorus and Phosphate in Peter and Paul Lakes")
```

```
## `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).
```



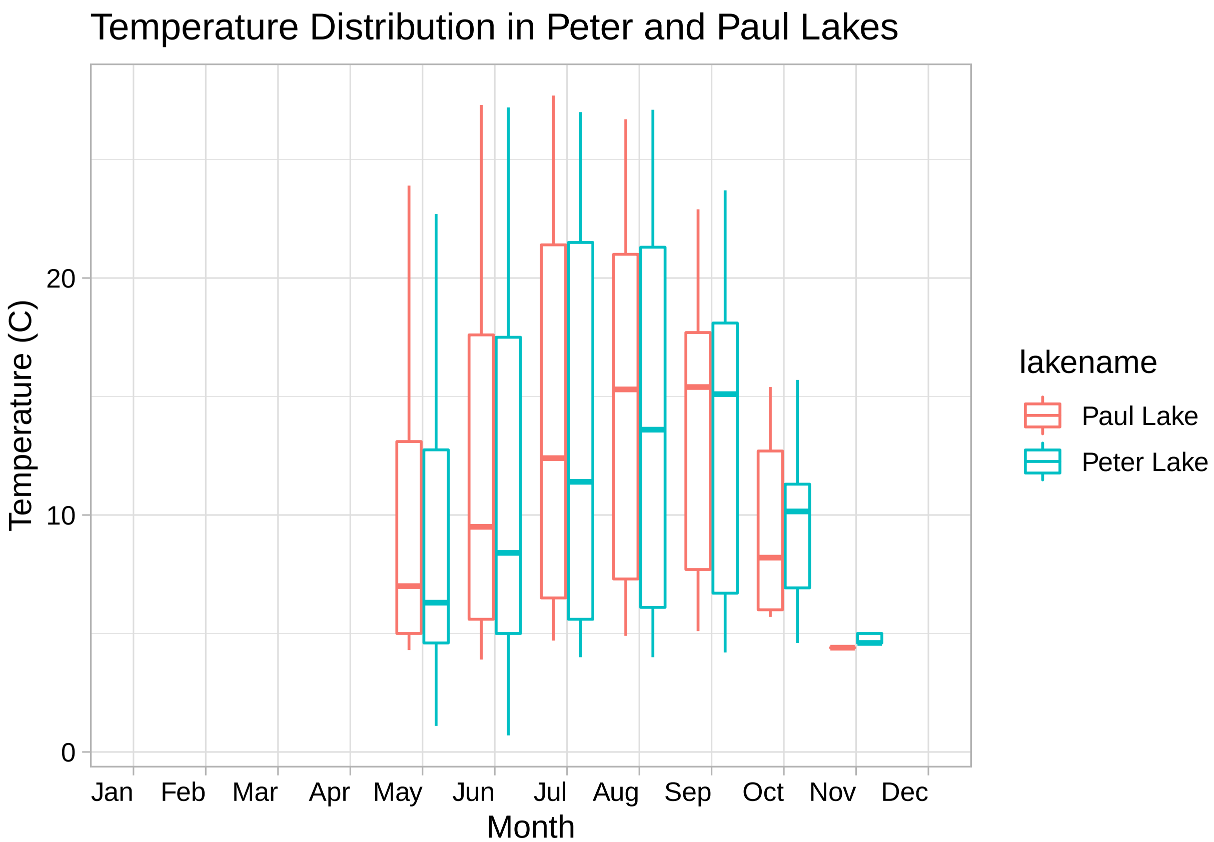
5. [NTL-LTER] Make three separate boxplots of (a) temperature, (b) TP, and

(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.

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

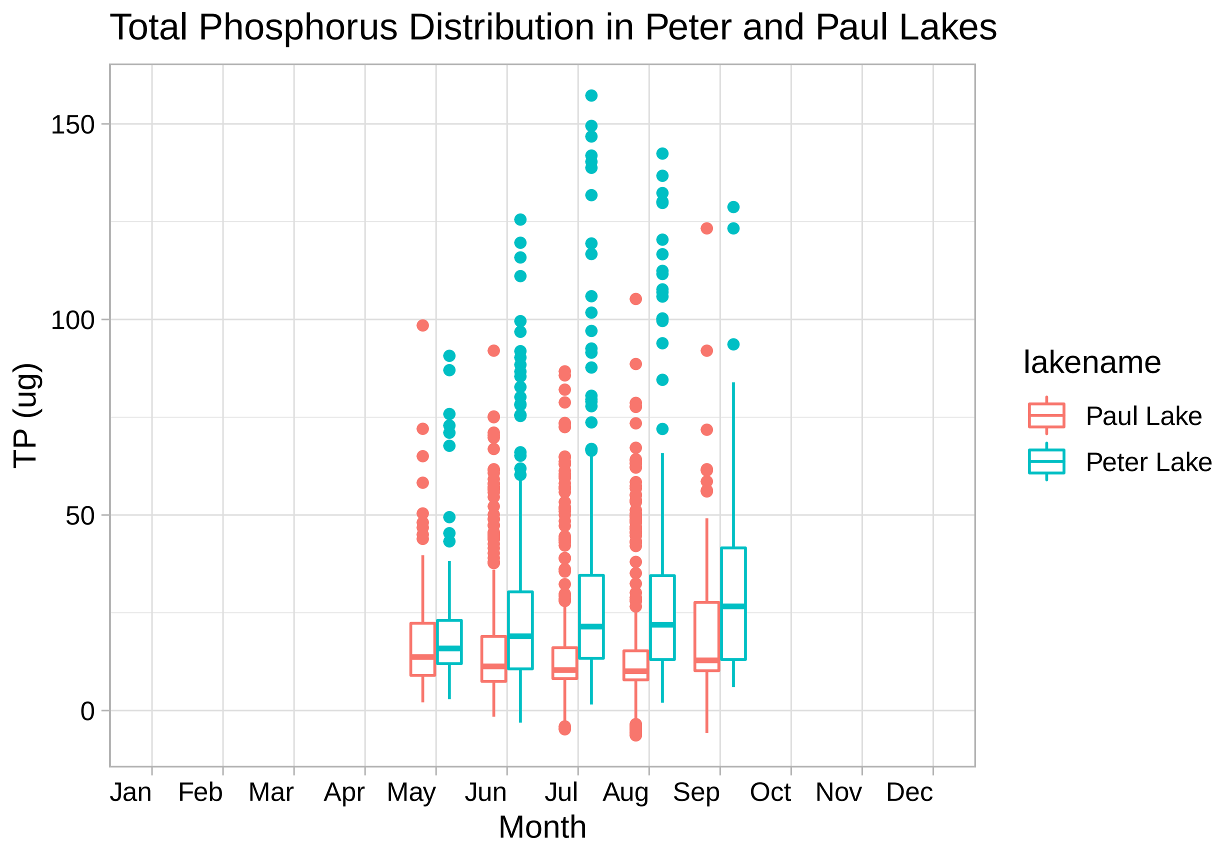
```
# 5
PeterPaul$month <- factor(PeterPaul$month,
    levels = c(1:12))

temperature_boxplot <- ggplot(PeterPaul,
    aes(x = month, y = temperature_C)) +
    geom_boxplot(aes(color = lakename)) +
    ylab("Temperature (C)") +
    xlab("Month") + scale_x_discrete(drop = FALSE,
    labels = month.abb) +
    theme(legend.position = "right",
        axis.text.x = element_text(hjust = 1)) +
    ggtitle("Temperature Distribution in Peter and Paul Lakes")
print(temperature_boxplot)
```

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



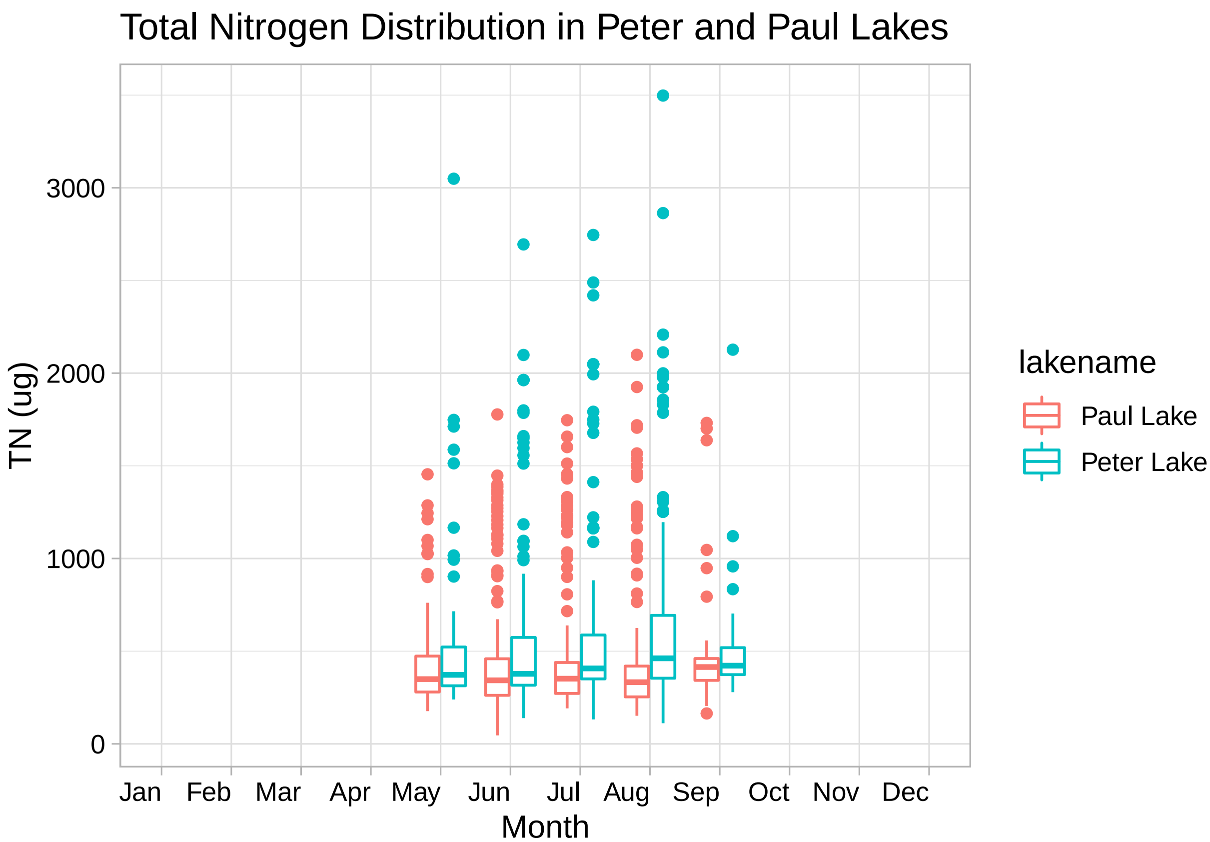

```
TP_boxplot <- ggplot(PeterPaul,
    aes(x = month, y = tp_ug)) +
    geom_boxplot(aes(color = lakename)) +
    ylab("TP (ug)") +
    xlab("Month") + scale_x_discrete(drop = FALSE,
    labels = month.abb) +
    theme(legend.position = "right",
        axis.text.x = element_text(hjust = 1)) +
    ggtitle("Total Phosphorus Distribution in Peter and Paul Lakes")
print(TP_boxplot)
```

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




```
TN_boxplot <- ggplot(PeterPaul,
    aes(x = month, y = tn_ug)) +
    geom_boxplot(aes(color = lakename)) +
    ylab("TN (ug)") +
    xlab("Month") + scale_x_discrete(drop = FALSE,
    labels = month.abb) +
    theme(legend.position = "right",
        axis.text.x = element_text(hjust = 1)) +
    ggtitle("Total Nitrogen Distribution in Peter and Paul Lakes")
print(TN_boxplot)
```

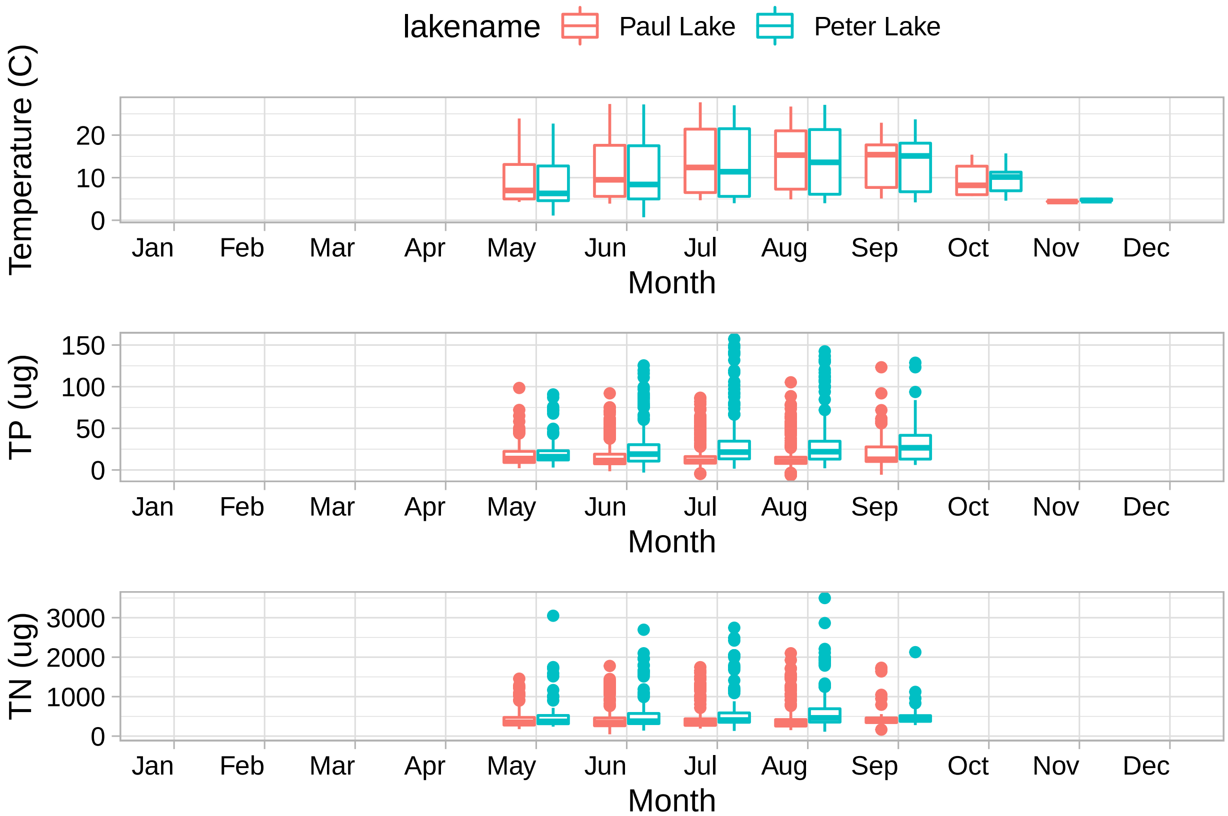
```
## Warning: Removed 21583 rows containing non-finite values (stat_boxplot).
```




```
cowplot <- plot_grid(temperature_boxplot +
    ggtitle(NULL) + theme(legend.position = "top"),
    TP_boxplot + theme(legend.position = "none") +
        ggtitle(NULL),
    TN_boxplot + theme(legend.position = "none") +
        ggtitle(NULL),
    nrow = 3, align = "v",
    rel_heights = c(12,
        9, 9))
```

```
## 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).
```

```
print(cowplot)
```

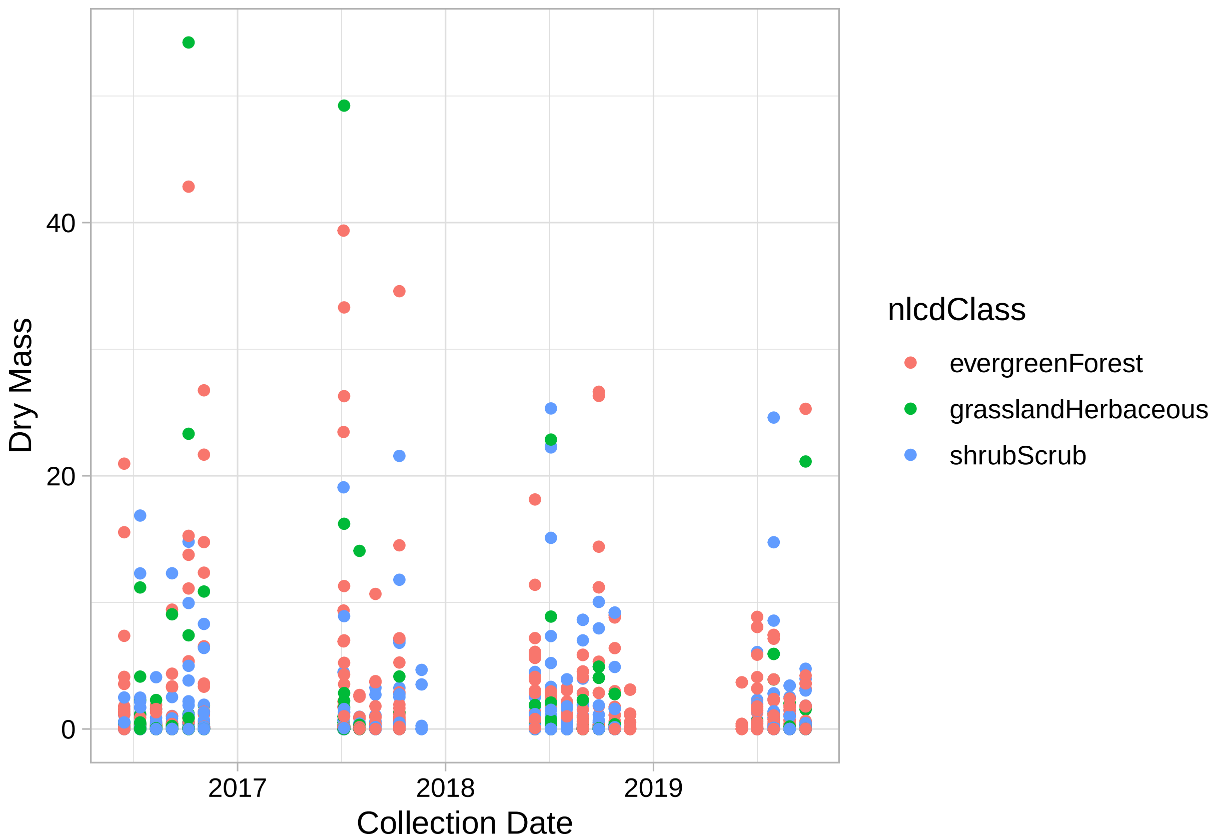


Question: What do you observe about the variables of interest over seasons and between lakes?

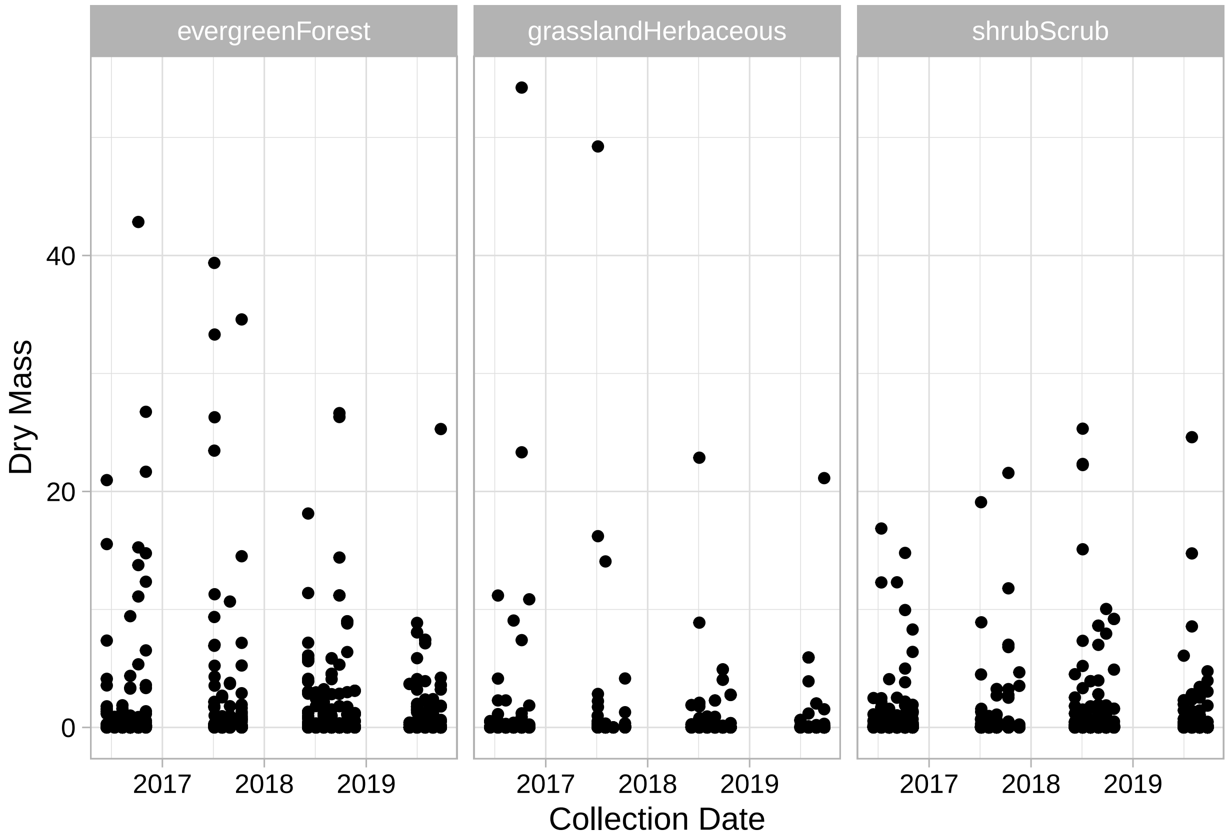
> Answer: The Temperatures lower and upper quartiles seem to be relatively similar through comparison of months between Paul and Peter lake with Peter Lake having the slightly lower median for the most part. The variability for Peter lake in terms of total phosphorus seems to be grater than that of Paul lake due to the range and outliers. The ranges seem to be similar for total nitrogen, however, it appears as though Peter Lake might contain larger quantities of Nitrogen based on their higher quartiles. In general, it seems as though Peter lake has more nutrient quantities with larger variabilities and the nutrients for peter lake increase through the months whilst the nutrient levels fluctuate by decreasing through july and august for Paul 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

ggplot(subset(NiwotRidge,
    functionalGroup = "Needles"),
    aes(x = collectDate,
        y = dryMass)) +
    assignment_theme +
    geom_point(aes(color = nlcdClass)) +
    ylab("Dry Mass") +
    xlab("Collection Date")
```




```
# 7
ggplot(subset(NiwotRidge,
    functionalGroup = "Needles"),
    aes(x = collectDate,
        y = dryMass)) +
    assignment_theme +
    geom_point() + ylab("Dry Mass") +
    xlab("Collection Date") +
    facet_wrap(~nlcdClass)
```



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

> Answer: Because these graphs are meant to compare dry mass in different classes, plot 6 is more effective as it allows us to more easily compare the distribution of dry mass in different years in relation to their classes. Plot 7 makes it difficult to as the years are spaced so far apart it is difficult to observe relative differences in dry mass that could have a potential correlation to class.