Assignment 5: Data Visualization

Logan Dye

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()
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   1.2.0      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
```

```
PeterPaulChemNut <- read.csv("./Data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv
    stringsAsFactors = TRUE)
LitterData <- read.csv("./Data/Processed/NEON_NIWO_Litter_mass_trap_Processed.csv",
    stringsAsFactors = TRUE)

# 2
str(LitterData)
```

```
## 'data.frame':    1692 obs. of  13 variables:
##  $ plotID          : Factor w/ 12 levels "NIWO_040","NIWO_041",..: 9 8 9 11 7 7 4 4 4 4 ...
##  $ trapID          : Factor w/ 15 levels "NIWO_040_139",..: 11 10 11 13 9 9 5 5 5 5 ...
##  $ collectDate     : Factor w/ 24 levels "2016-06-16","2016-07-14",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ functionalGroup : Factor w/ 8 levels "Flowers","Leaves",..: 6 5 8 6 4 2 2 6 7 8 ...
##  $ dryMass         : num  0 0.27 0.12 0 1.11 0 0 0 0.07 0.02 ...
##  $ qaDryMass       : Factor w/ 2 levels "N","Y": 1 1 1 1 2 1 1 1 1 1 ...
##  $ subplotID       : int  31 41 31 32 32 32 40 40 40 40 ...
##  $ decimalLatitude : num  40.1 40 40.1 40 40 ...
##  $ decimalLongitude: num  -106 -106 -106 -106 -106 ...
##  $ elevation       : num  3477 3413 3477 3373 3446 ...
##  $ nlcdClass       : Factor w/ 3 levels "evergreenForest",..: 3 1 3 1 3 3 2 2 2 2 ...
##  $ plotType        : Factor w/ 1 level "tower": 1 1 1 1 1 1 1 1 1 1 ...
##  $ geodeticDatum   : Factor w/ 1 level "WGS84": 1 1 1 1 1 1 1 1 1 1 ...
```

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

str(PeterPaulChemNut)
```

```
## 'data.frame':    23008 obs. of  15 variables:
##  $ lakename       : Factor w/ 2 levels "Paul Lake","Peter Lake": 1 1 1 1 1 1 1 1 1 1 ...
##  $ year4          : int  1984 1984 1984 1984 1984 1984 1984 1984 1984 1984 ...
##  $ daynum         : int  148 148 148 148 148 148 148 148 148 148 ...
##  $ month          : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ sampledate     : Factor w/ 1103 levels "1984-05-27","1984-05-28",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ depth          : num  0 0.25 0.5 0.75 1 1.5 2 3 4 5 ...
##  $ temperature_C  : num  14.5 NA NA NA 14.5 NA 14.2 11 7 6.1 ...
##  $ dissolvedOxygen: num  9.5 NA NA NA 8.8 NA 8.6 11.5 11.9 2.5 ...
##  $ irradianceWater: num  1750 1550 1150 975 870 610 420 220 100 34 ...
##  $ irradianceDeck : num  1620 1620 1620 1620 1620 1620 1620 1620 1620 1620 ...
##  $ tn_ug          : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ tp_ug          : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ nh34           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ no23           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ po4            : num  NA NA NA NA NA NA NA NA NA NA ...
```

```
PeterPaulChemNut$year4 <- as.Date(PeterPaulChemNut$sampledate, format = "%Y-%m-%d")
```

Define your theme

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

```
# 3
mytheme <- theme_classic(base_size = 14) + theme(axis.text = element_text(color = "red"),
    legend.position = "right")
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.

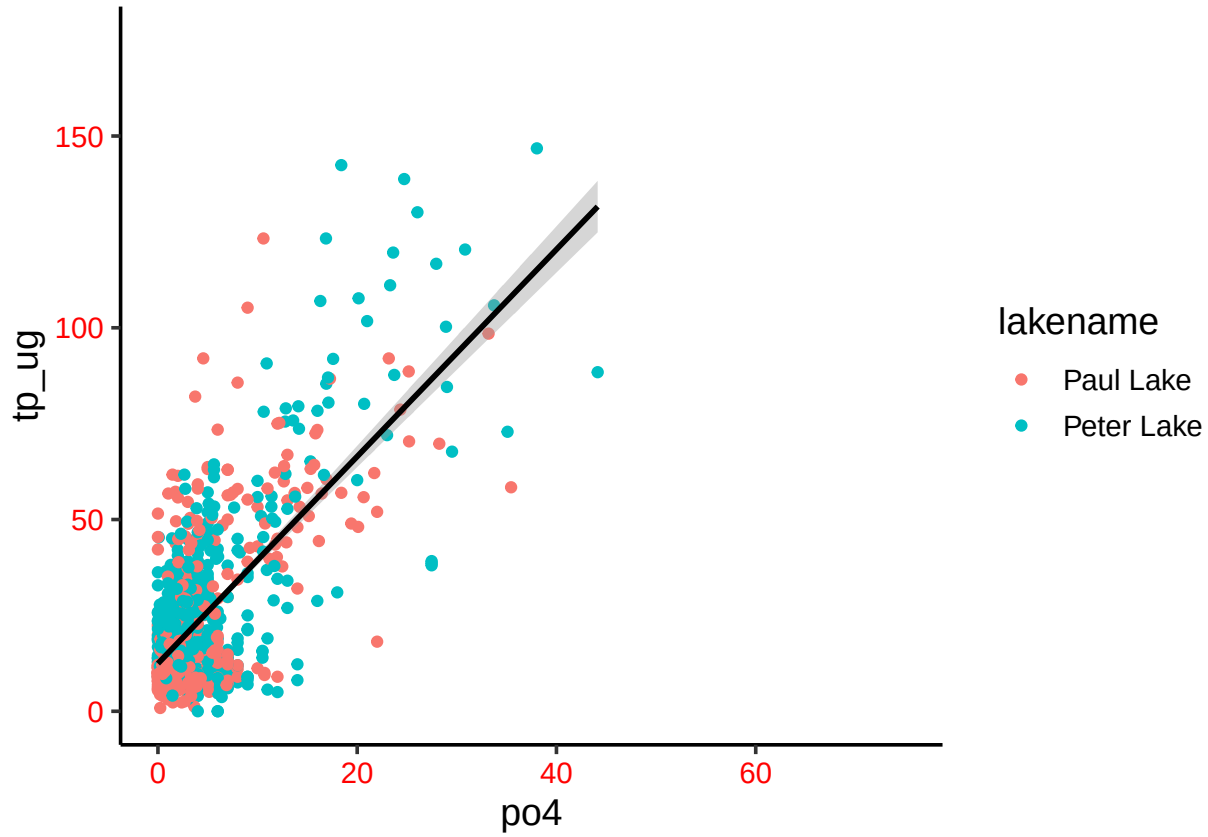
4. [NTL-LTER] Plot total phosphorus (`tp_ug`) by phosphate (`po4`), with separate aesthetics for Peter and Paul lakes. Add a line of best fit and color it black. Adjust your axes to hide extreme values (hint: change the limits using `xlim()` and/or `ylim()`).

```
# 4

PhosphorusByPhosphate <- ggplot(PeterPaulChemNut, aes(x = po4, y = tp_ug)) + geom_point(aes(color = lak
    geom_smooth(method = "lm", color = "black") + xlim(0, 75) + ylim(0, 175)
print(PhosphorusByPhosphate)
```

```
## `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).
```



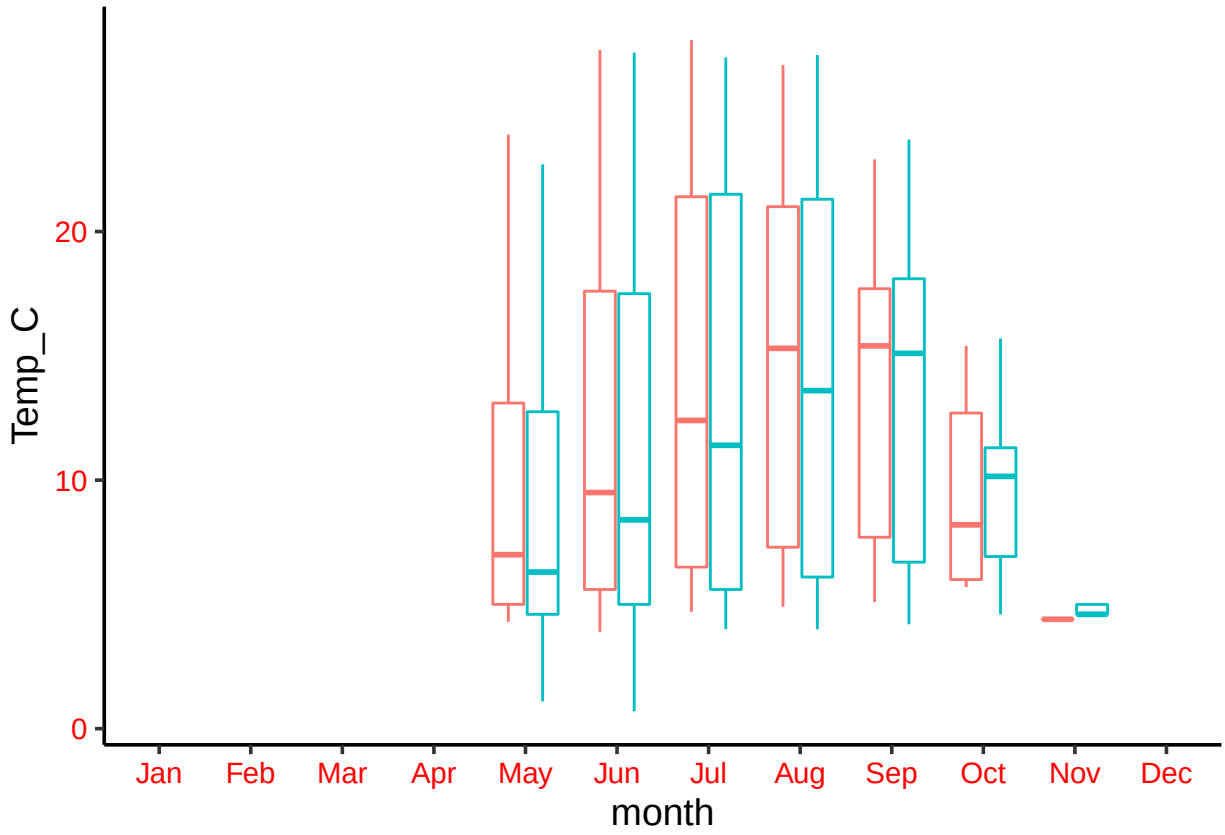
5. [NTL-LTER] Make three separate boxplots of (a) temperature, (b) TP, and

(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

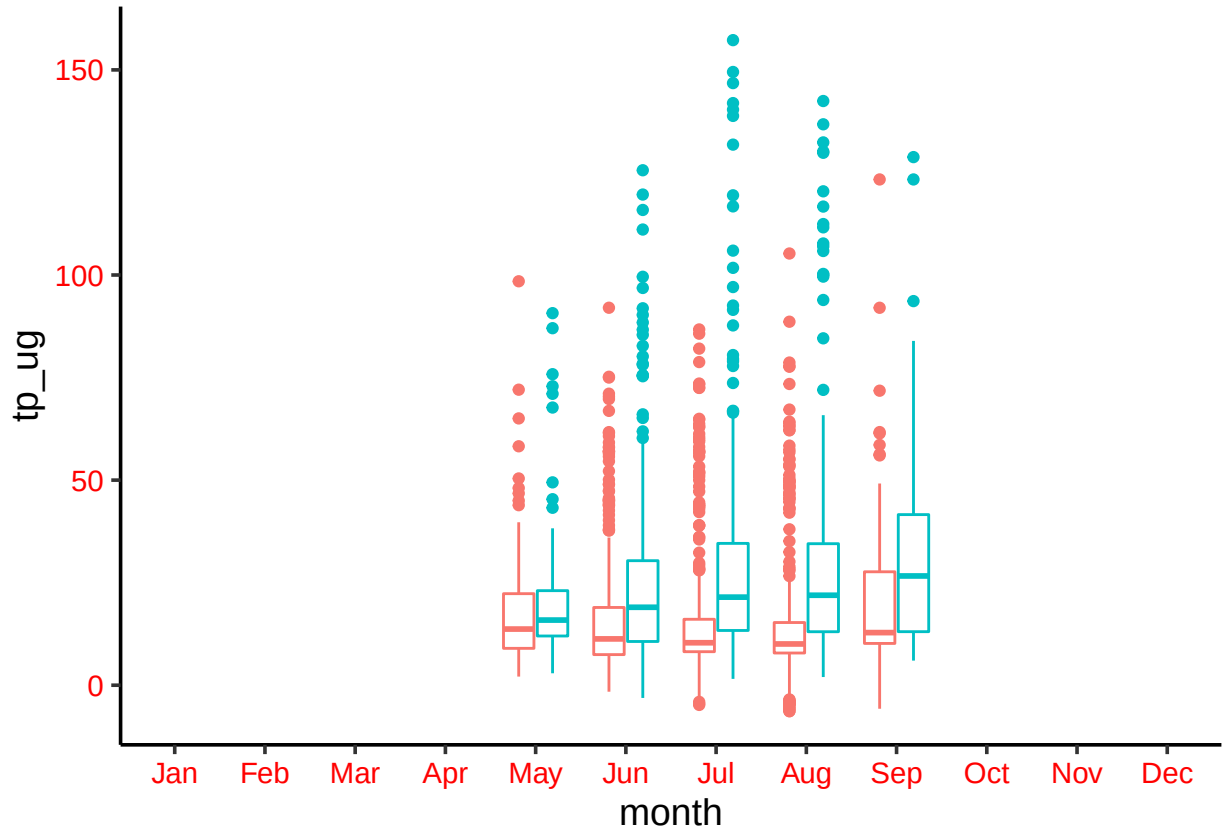
```
# 5
PeterPaulChemNut$month <- factor(PeterPaulChemNut$month, levels = c(1:12))

TempBPlot <- ggplot(PeterPaulChemNut, aes(x = month, y = temperature_C)) + geom_boxplot(aes(color = lak
    scale_x_discrete(label = month.abb, drop = FALSE) + ylab("Temp_C") + theme(legend.position = "none"
print(TempBPlot)
```

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



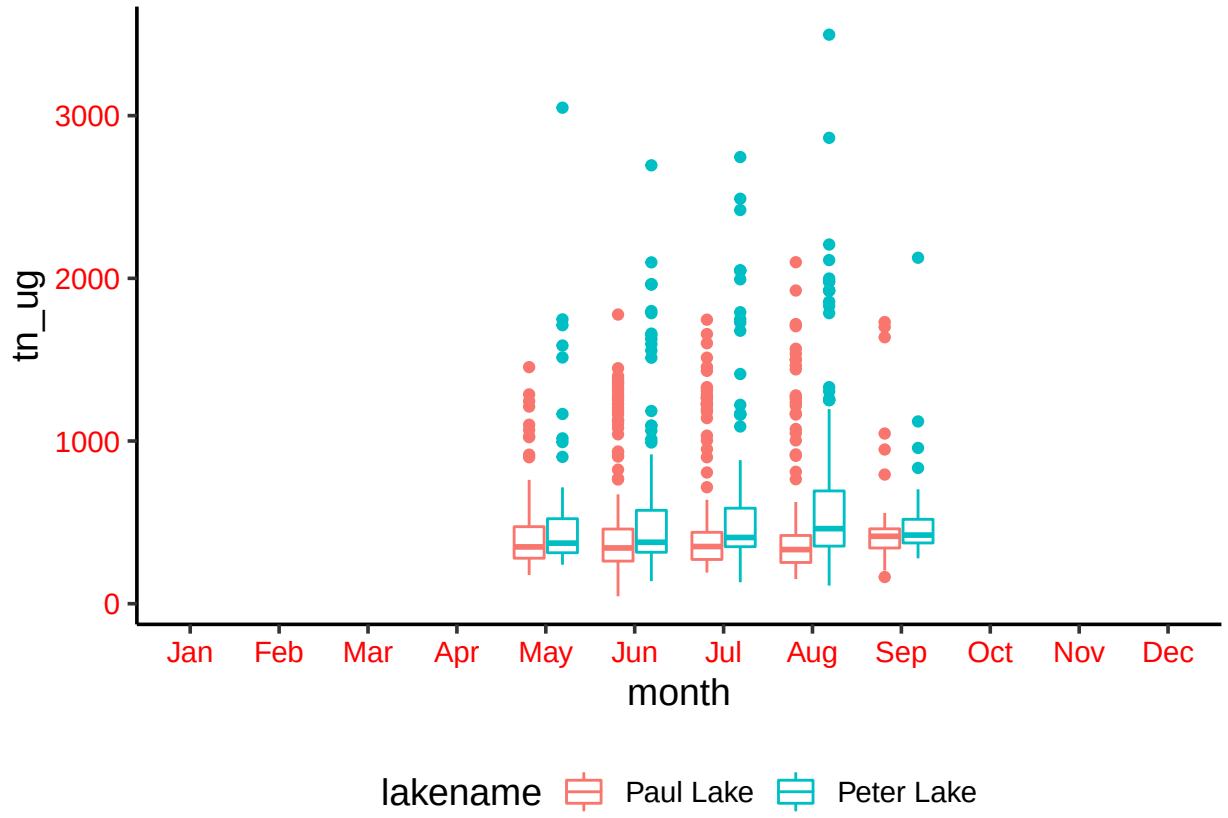

```
TPBPlot <- ggplot(PeterPaulChemNut, aes(x = month, y = tp_ug)) + geom_boxplot(aes(color = lakename)) +
    scale_x_discrete(label = month.abb, drop = FALSE) + theme(legend.position = "none")
print(TPBPlot)
```

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




```
TNBPlot <- ggplot(PeterPaulChemNut, aes(x = month, y = tn_ug)) + geom_boxplot(aes(color = lakename)) +
    scale_x_discrete(label = month.abb, drop = FALSE) + theme(legend.position = "bottom")
print(TNBPlot)
```

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

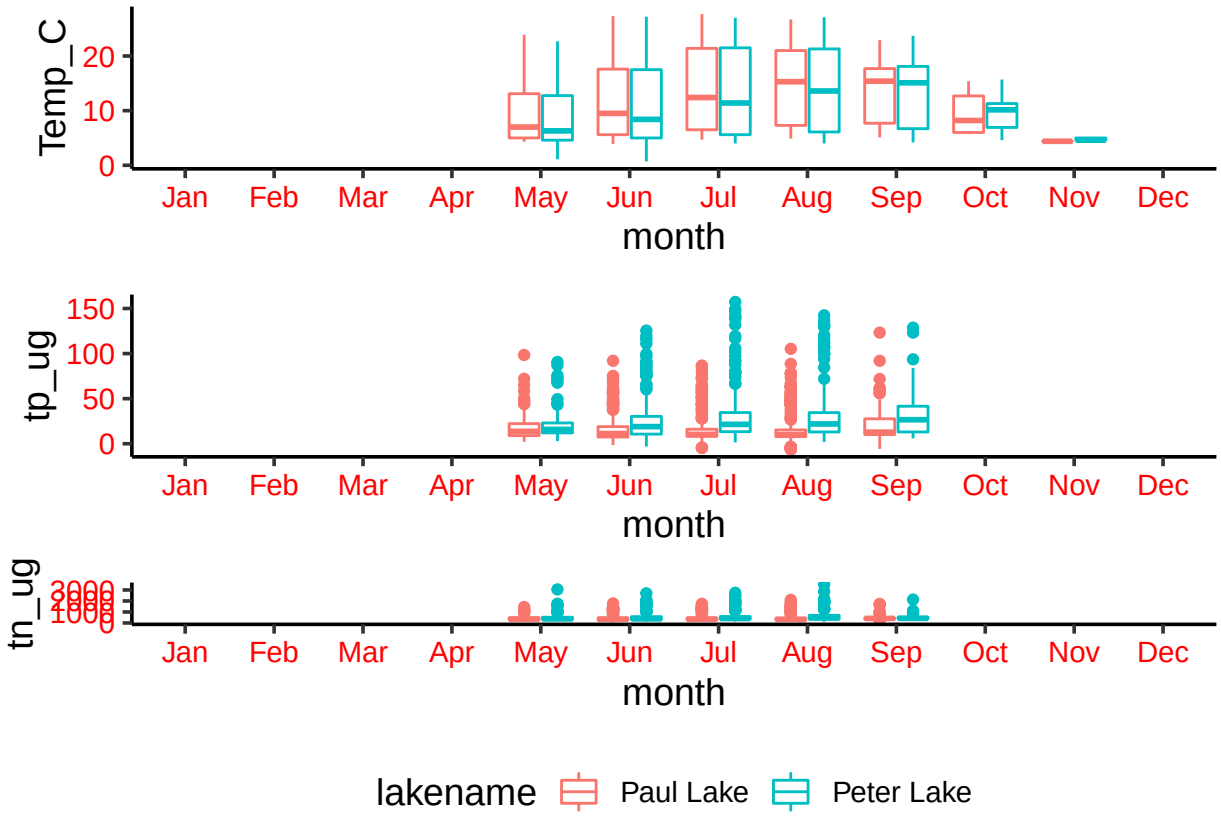



```
plot_grid(TempBPlot, TPBPlot, TNBPlot, nrow = 3, align = "v", axis = c("l"))
```

```
## 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).
```



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

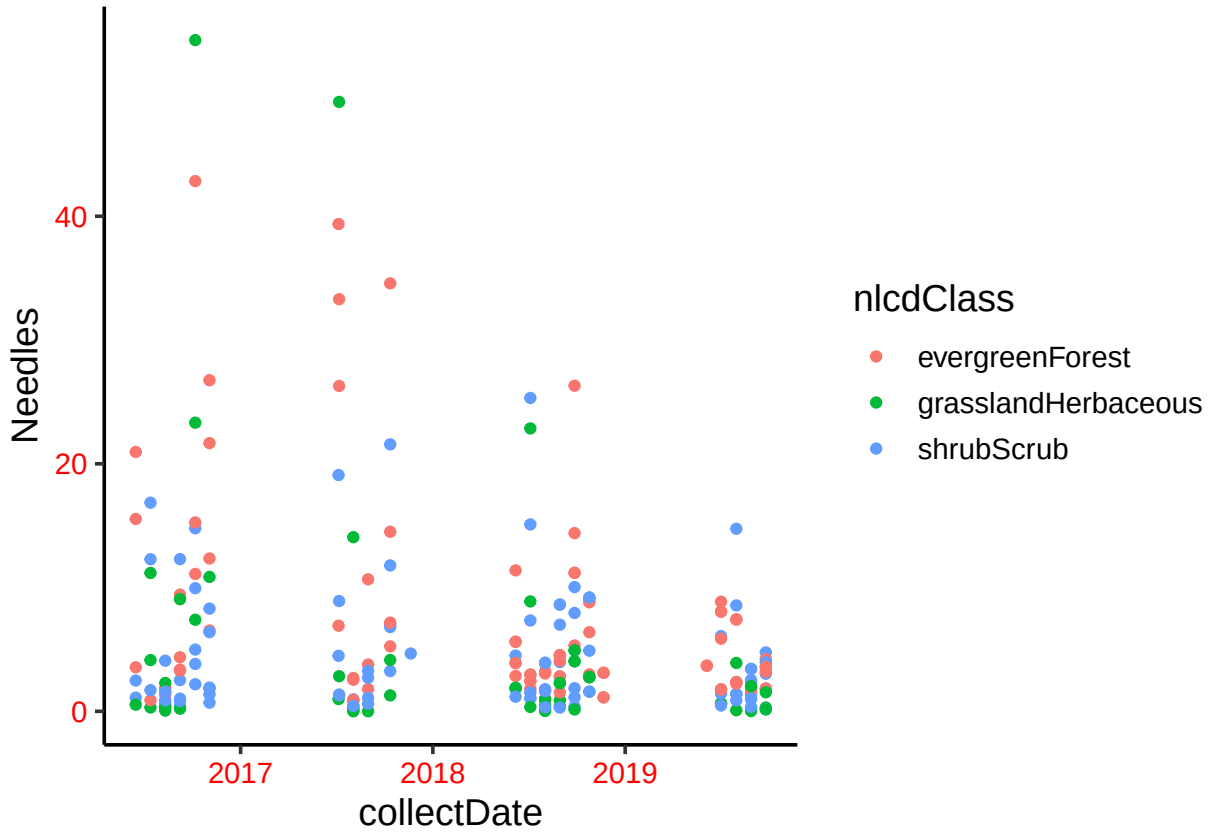
> Answer: The temperatures between the two lakes move in relative synchronicity. Summer months have larger medians. This makes sense condidering their geographic relation to eachother and the fact that summer months are hotter. For TP, Peter lake consistantly has higher values than Paul Lake. TP for Peter lake consistantly rises from May through September, while Paul lake stays at relativly the same level. For TN, the medians are similar with Peter lake being slightly higher, but not by much. The difference between the lakes reveals itself in the high value outliers. Peter lake has upper level outliers that are of much greater value than Paul lake in every month. There is not a dramatic difference in TN seasonally.

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.

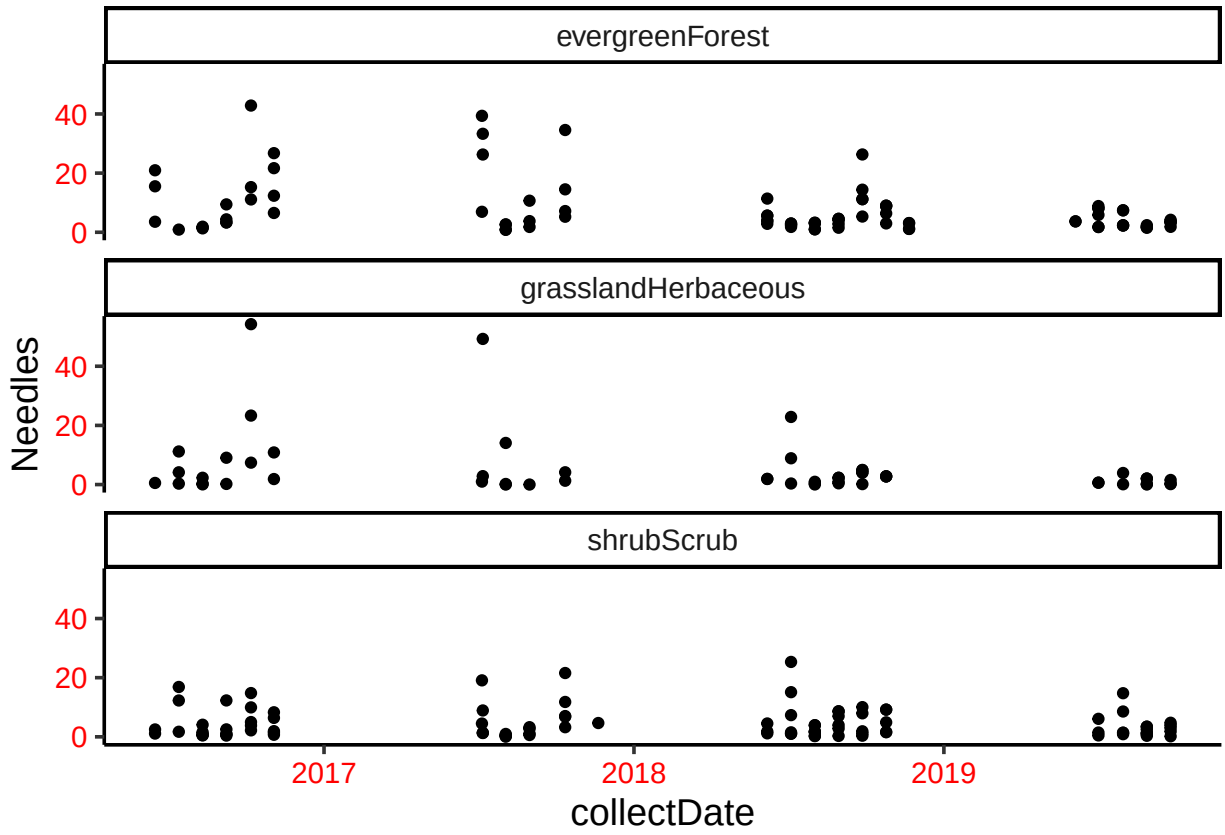
```
# 6
LitterDataSpreadNeedles <- pivot_wider(LitterData, names_from = functionalGroup,
    values_from = dryMass)

NeedleMassPlot <- ggplot(LitterDataSpreadNeedles) + geom_point(aes(x = collectDate,
    y = Needles, color = nlcdClass))
print(NeedleMassPlot)
```

```
## Warning: Removed 108 rows containing missing values (geom_point).
```




```
# 7
NeedleMassFacetPlot <- ggplot(LitterDataSpreadNeedles) + geom_point(aes(x = collectDate,
    y = Needles)) + facet_wrap(vars(nlcdClass), nrow = 3)
print(NeedleMassFacetPlot)
```

```
## Warning: Removed 108 rows containing missing values (geom_point).
```



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

Answer: I think the faceted plot is more effective, because when all of the data points are overlapping with eachother it looks clutterd and it is difficult to view the differences between the three different colors. When there are facets it is easier to view the differences between the three different classes.