Assignment 5: Data Visualization

Nicole Eastman

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 Set-up
setwd("/home/guest/R/EDA Fall 2022")
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
```

```
library(formatR)
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
PP.Nutrients <- read.csv("Data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv")
Niwot.Litter <- read.csv("Data/Processed/NEON_NIWO_Litter_mass_trap_Processed.csv")

# 2 Date Format
PP.Nutrients$sampledate <- as.Date(PP.Nutrients$sampledate, "%Y-%m-%d")
Niwot.Litter$collectDate <- as.Date(Niwot.Litter$collectDate, "%Y-%m-%d")
```

Define your theme

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

```
# 3 Theme
mytheme <- theme_classic(base_size = 12) + 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.

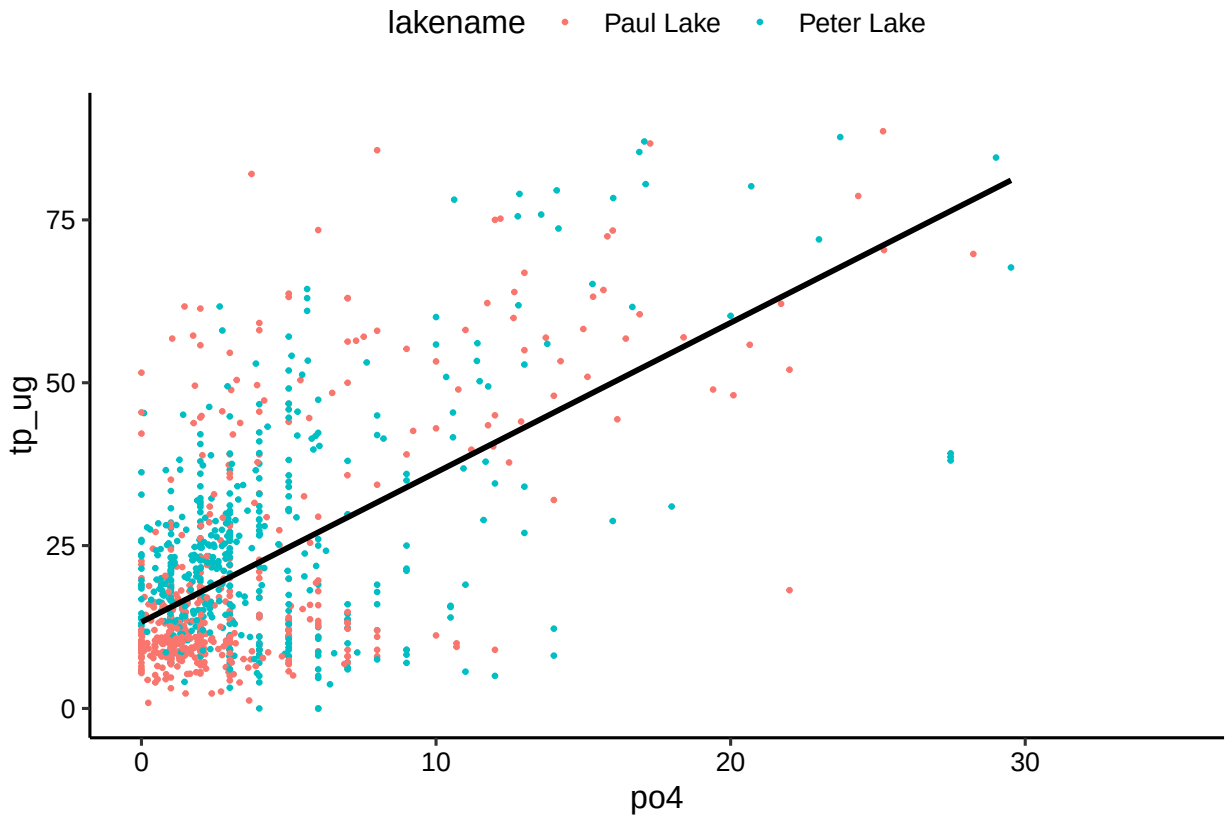
4. [NTL-LTER] Plot total phosphorus (`tp_ug`) by phosphate (`po4`), with separate aesthetics for Peter and Paul lakes. Add a line of best fit and color it black. Adjust your axes to hide extreme values (hint: change the limits using `xlim()` and/or `ylim()`).

```
# 4 Scatterplot Construction
Nutrients.Scatter <- ggplot(PP.Nutrients, aes(x = po4, y = tp_ug)) + geom_point(aes(color = lakename),
    size = 0.5) + xlim(0, 35) + ylim(0, 90) + geom_smooth(method = lm, se = FALSE,
    color = "black")
print(Nutrients.Scatter)
```

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

```
## Warning: Removed 21972 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 21972 rows containing missing values (geom_point).
```



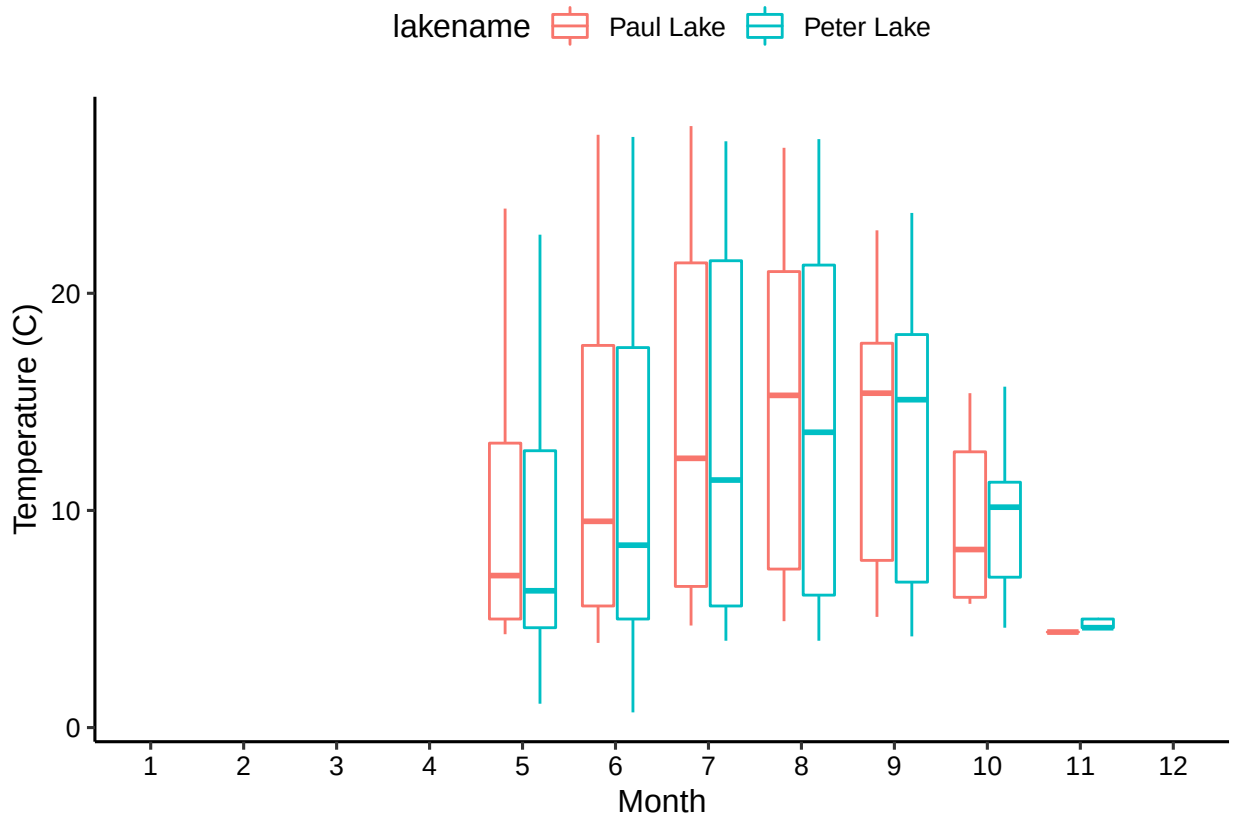
5. [NTL-LTER] Make three separate boxplots of (a) temperature, (b) TP, and

(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

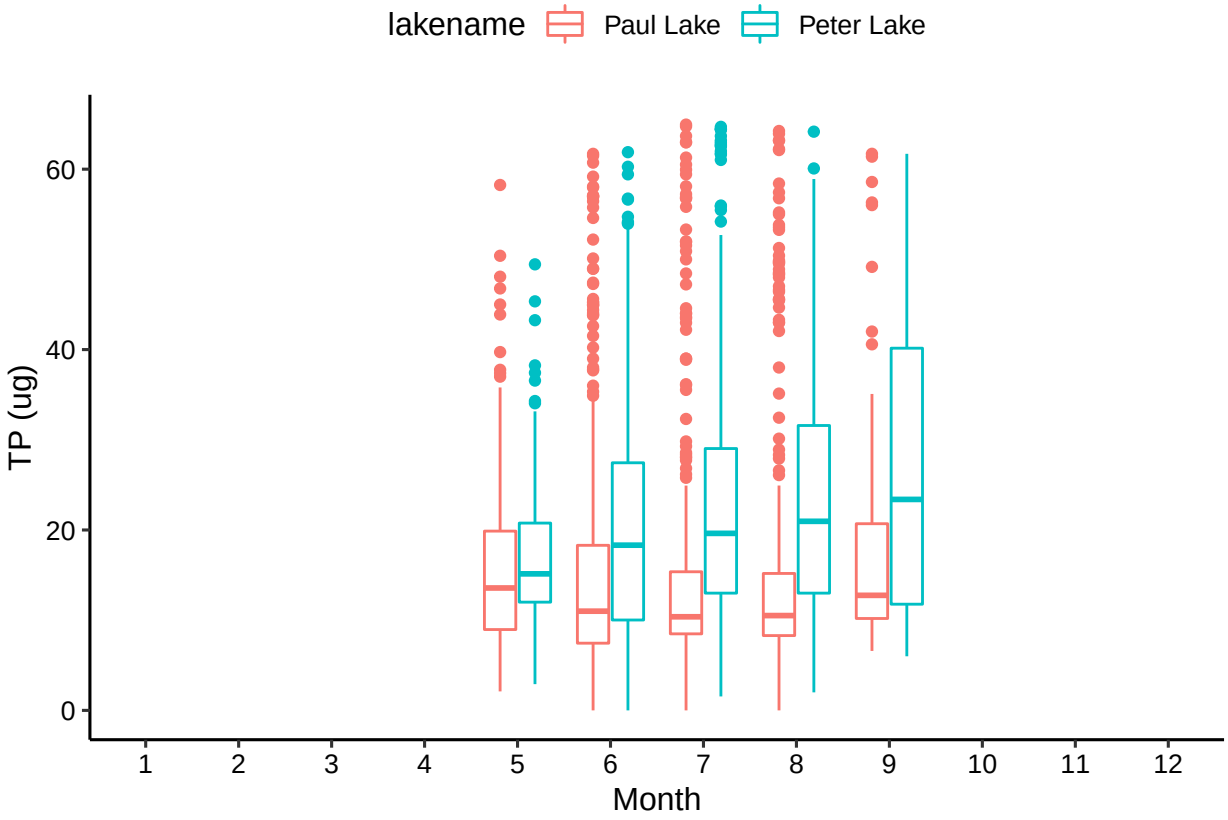
```
# 5a Boxplot Construction
PP.Nutrients$month <- factor(PP.Nutrients$month, levels = c(1:12))
Box.Temp <- ggplot(PP.Nutrients, aes(x = month, y = temperature_C, color = lakename)) +
    geom_boxplot() + scale_x_discrete(drop = FALSE) + labs(x = "Month", y = "Temperature (C)")
print(Box.Temp)
```

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




```
# 5b
Box.TP <- ggplot(PP.Nutrients, aes(x = month, y = tp_ug)) + geom_boxplot(aes(color = lakename)) +
    scale_x_discrete(drop = FALSE) + labs(x = "Month", y = "TP (ug)") + ylim(0, 65)
print(Box.TP)
```

```
## Warning: Removed 20870 rows containing non-finite values (stat_boxplot).
```

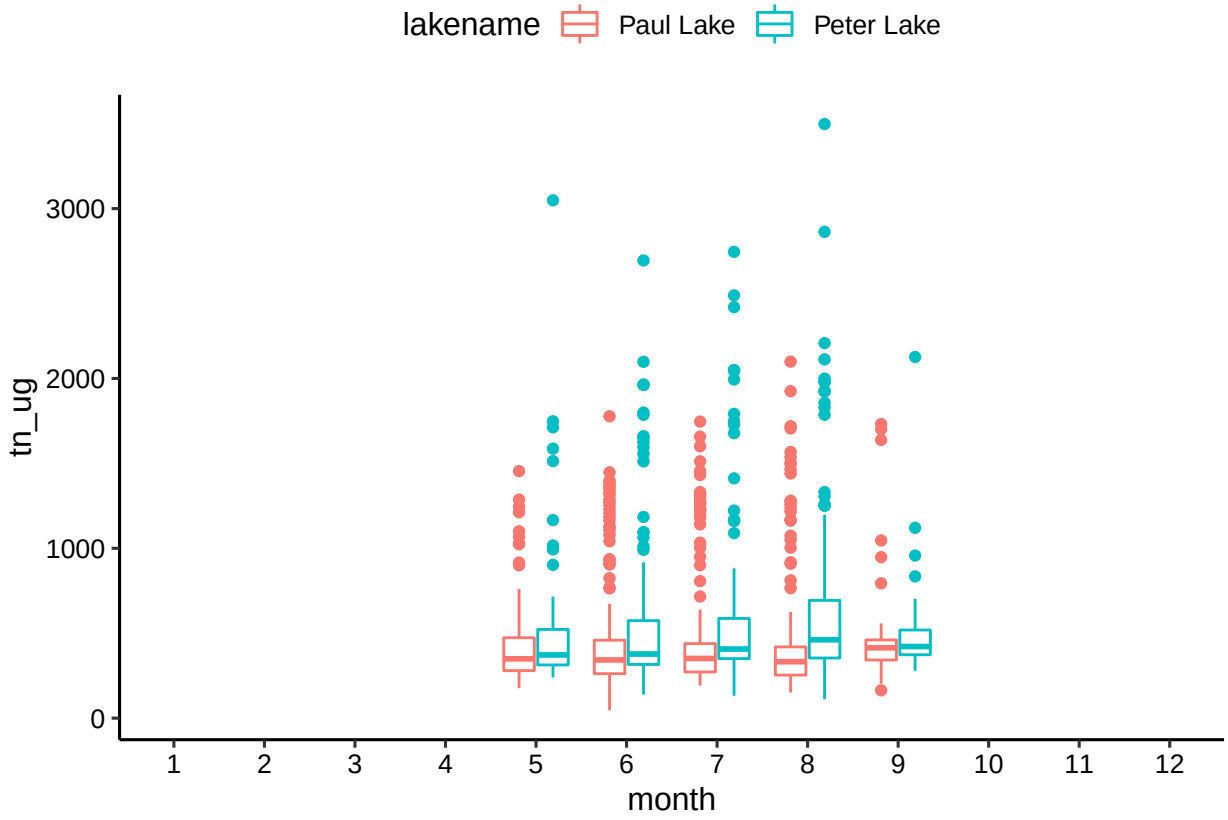



```
# 5c
Box.TN <- ggplot(PP.Nutrients, aes(x = month, y = tn_ug)) + geom_boxplot(aes(color = lakename)) +
    scale_x_discrete(drop = FALSE)
labs(x = "Month", y = "TN (ug)") + ylim(0, 780)
```

```
## NULL
```

```
print(Box.TN)
```

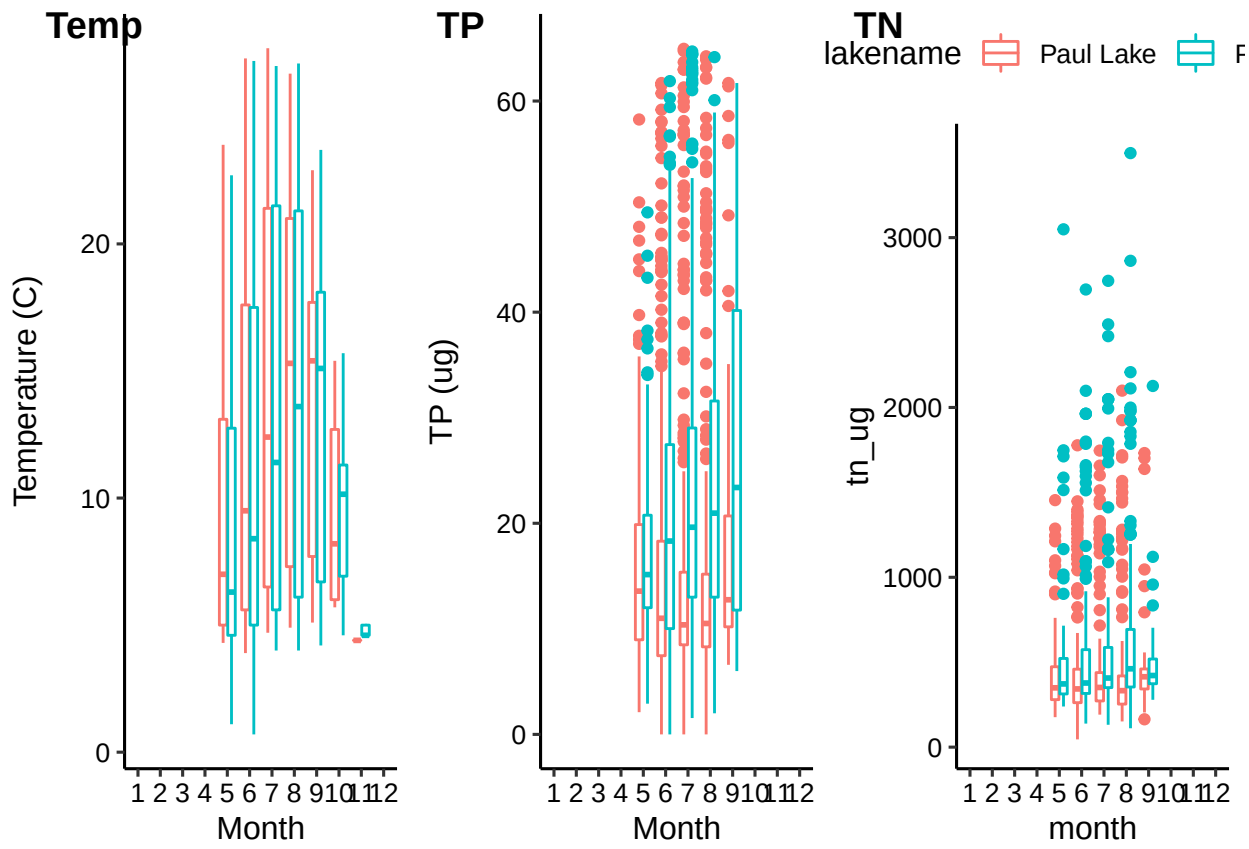
```
## Warning: Removed 21583 rows containing non-finite values (stat_boxplot).
```




```
# 5d Cowplot
Nutrients.Cowplot <- plot_grid(Box.Temp + theme(legend.position = "none"), Box.TP +
    theme(legend.position = "none"), Box.TN, nrow = 1, align = "vh", axis = "b",
    labels = c("Temp", "TP", "TN"))
```

```
## Warning: Removed 3566 rows containing non-finite values (stat_boxplot).
## Warning: Removed 20870 rows containing non-finite values (stat_boxplot).
## Warning: Removed 21583 rows containing non-finite values (stat_boxplot).
```

```
print(Nutrients.Cowplot)
```



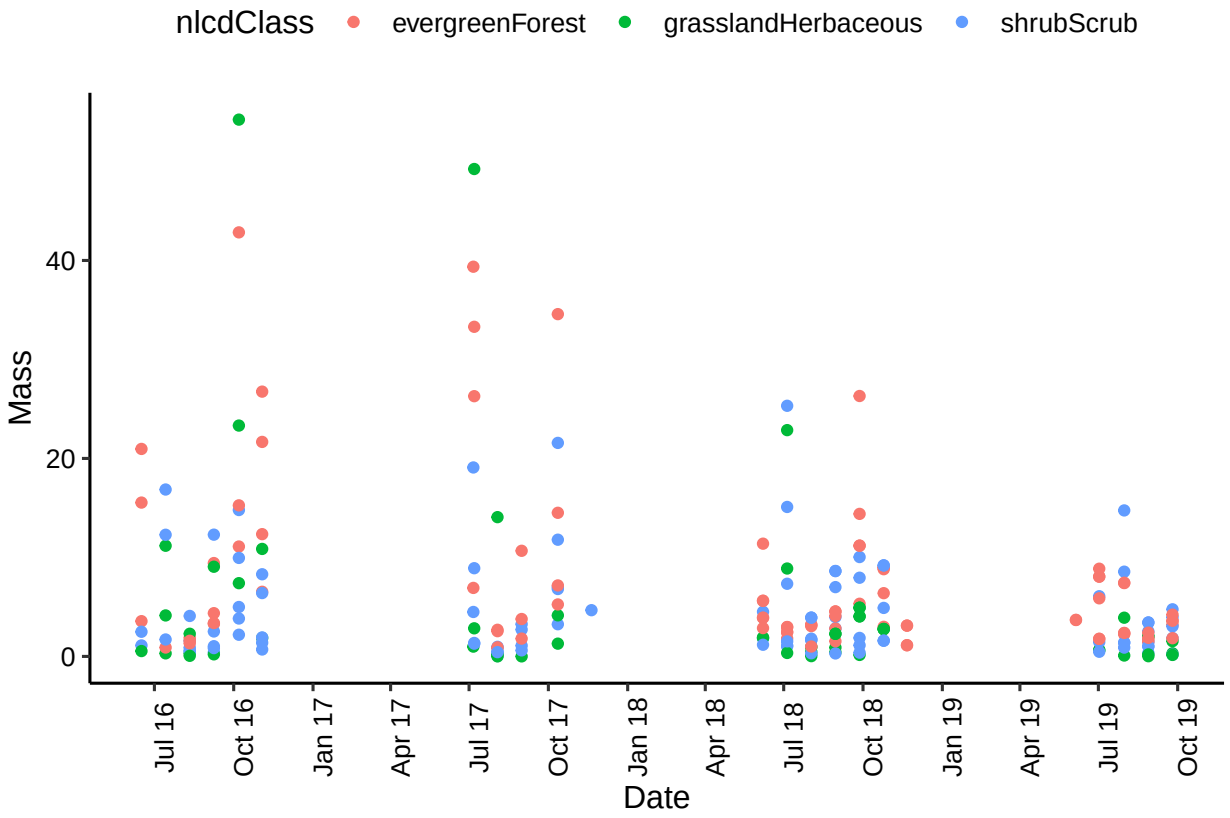
Question: What do you observe about the variables of interest over seasons and between lakes?

> Answer: The temperature boxplot indicates higher median temperatures in the summer months, and Paul Lake generally has higher median temperatures throughout the year. However, Peter Lake remains higher after temperatures start dropping in Paul Lake. Peter Lake has greater TP levels across the spring and summer months, and Paul lake has the greatest amount of high outliers. The TN boxplot indicates similar median TN values across the spring and summer seasons between lakes and months. However, there are a substantial amount of high lying outliers for both lakes.

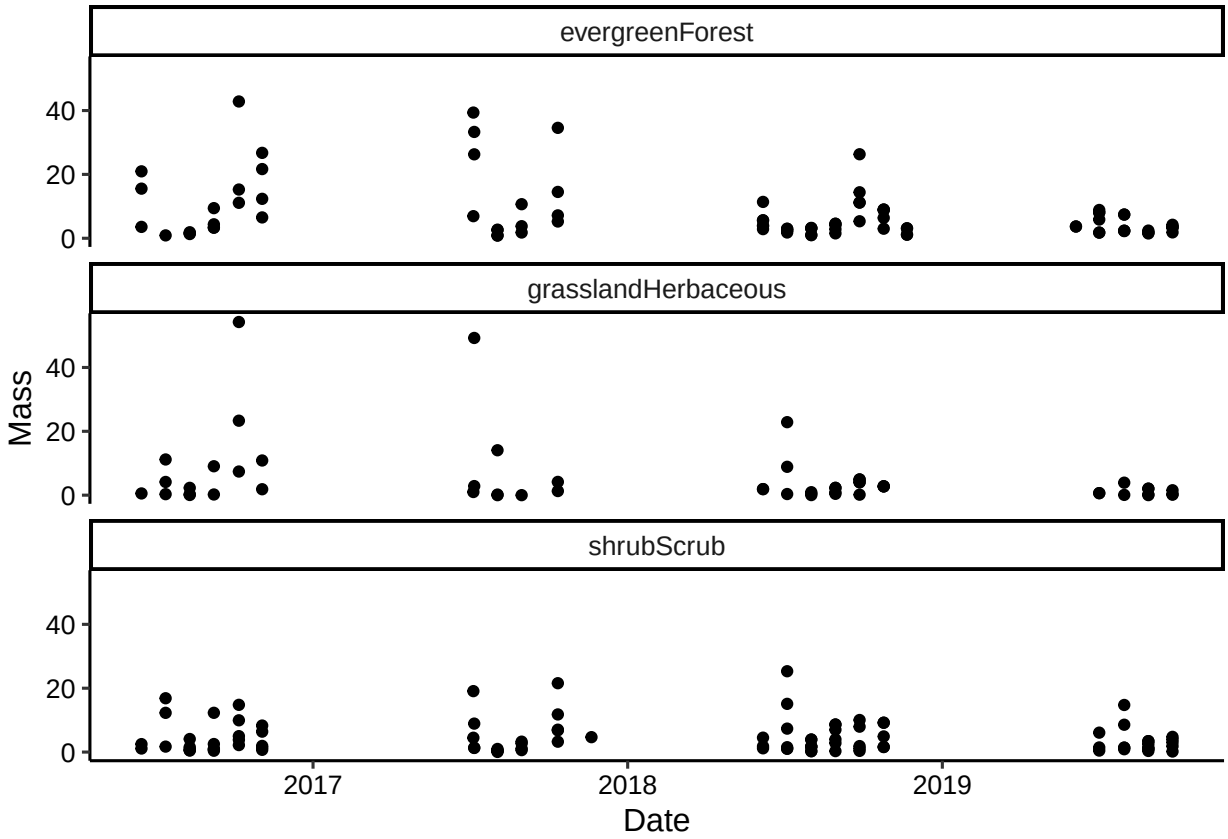
6. [Niwot Ridge] Plot a subset of the litter dataset by displaying only the "Needles" functional group. Plot the dry mass of needle litter by date and separate by NLCD class with a color aesthetic. (no need to adjust the name of each land use)

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

```
# 6 Subset of Litter Dataset: 'Needles'
Needles.Plot <- ggplot(filter(Niwot.Litter, functionalGroup == "Needles"), aes(x = collectDate,
    y = dryMass, color = nlcdClass)) + geom_point() + labs(x = "Date", y = "Mass") +
    scale_x_date(limits = as.Date(c("2016-06-16", "2019-09-25")), date_breaks = "3 months",
        date_labels = "%b %y") + theme(axis.text.x = element_text(angle = 90))
print(Needles.Plot)
```




```
# 7
Needles.faceted.plot <- ggplot(filter(Niwot.Litter, functionalGroup == "Needles"),
    aes(x = collectDate, y = dryMass)) + geom_point() + facet_wrap(vars(nlcdClass),
    nrow = 3) + labs(x = "Date", y = "Mass")
print(Needles.faceted.plot)
```



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

Answer: I think the faceted graph is more effective because it is easier to visualize the different mass values over the years for each given category when they are separated. It is difficult to see trends with the "busy" colored dots in #6.