Assignment 5: Data Visualization

Zhiyuan Chen

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_PeterP`
version) and the processed data file for the Niwot Ridge litter dataset (use the [`NEON_NIWO_Litter_mass_trap_Processe`
version).

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

```
# 1
getwd()
```

```
## [1] "D:/Rfiles/EDA-Fall2022"
```

```
# install.packages('tidyverse')
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()
```

```
# install.packages('lubridate')
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
# install.packages('cowplot')
library(cowplot)
```

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

```
nutrient_data <- read.csv("./Data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv",
    stringsAsFactors = TRUE)
litter_data <- read.csv("./Data/Processed/NEON_NIWO_Litter_mass_trap_Processed.csv",
    stringsAsFactors = TRUE)
# 2
nutrient_data$sampledate <- as.Date(nutrient_data$sampledate)
# nutrient_data
class(nutrient_data$sampledate)
```

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

```
litter_data$collectDate <- as.Date(litter_data$collectDate)
# litter_data
class(litter_data$collectDate)
```

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

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 = "black"),
    legend.position = "top")
theme_set(mytheme)
```

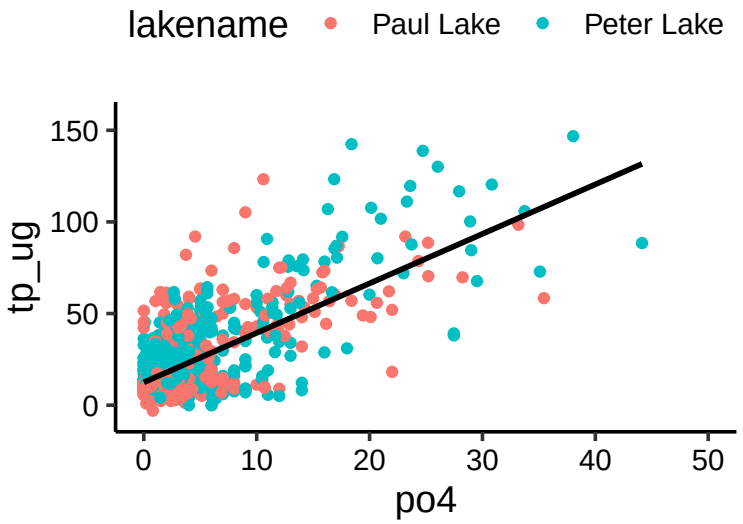
Create graphs

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
PeterPaulplot.Ex4 <- ggplot(nutrient_data, aes(x = po4, y = tp_ug, color = lakename)) +
    geom_point() + xlim(0, 50) + geom_smooth(method = "lm", se = FALSE, color = "black")
print(PeterPaulplot.Ex4)
```

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

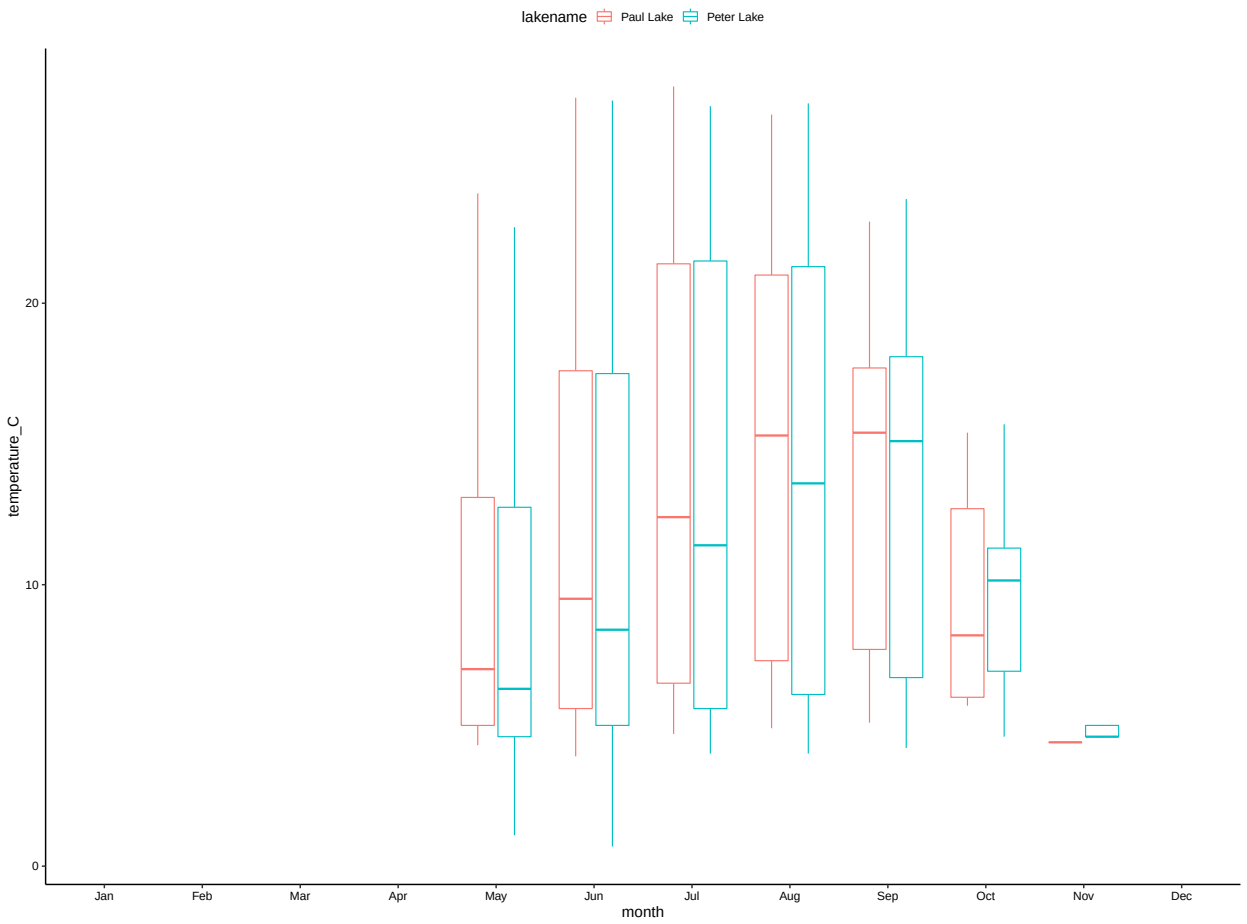


5. [NTL-LTER] Make three separate boxplots of (a) temperature, (b) TP, and

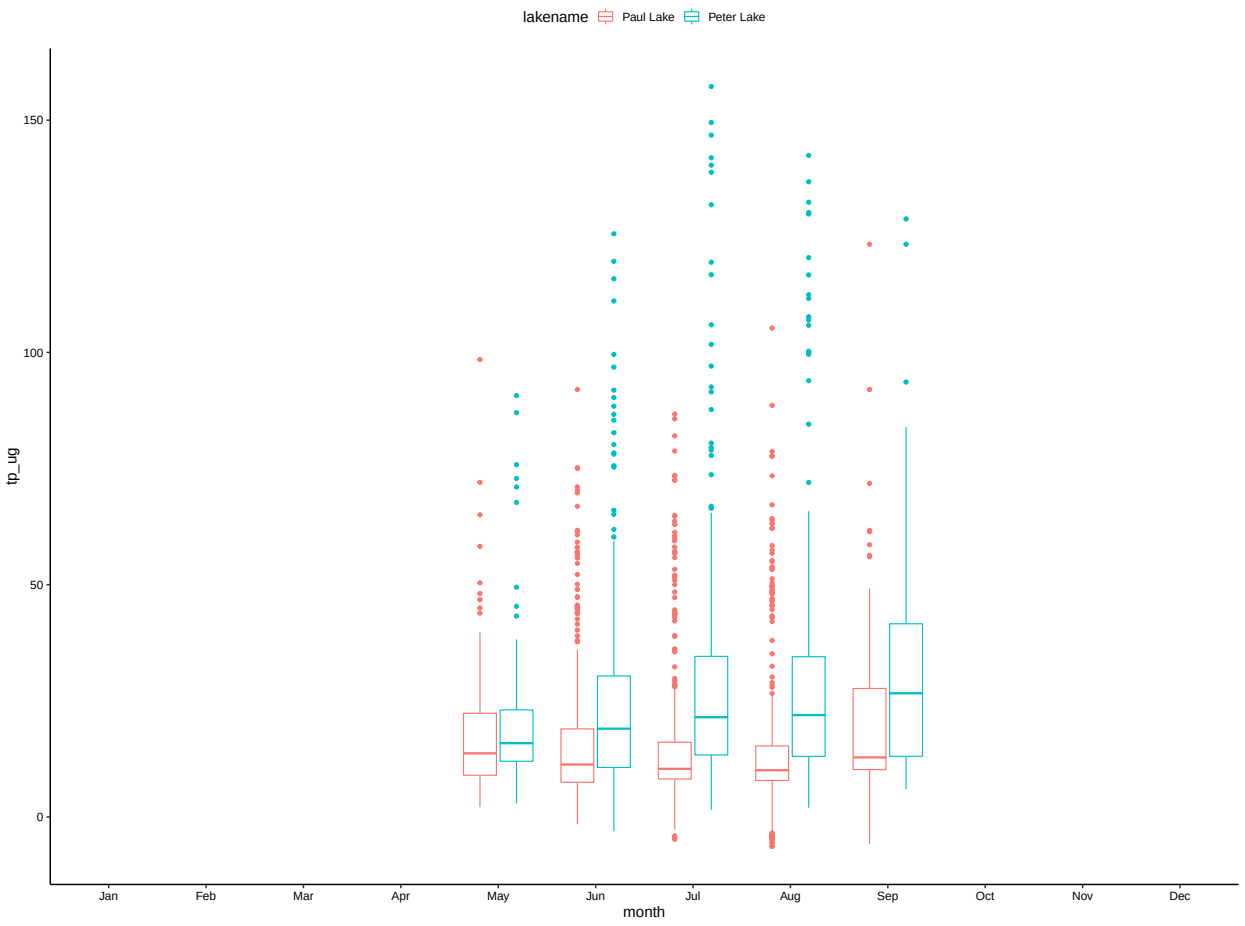
(c) TN, with month as the x axis and lake as a color aesthetic. Then, create a cowplot that combines the three graphs. Make sure that only one legend is present and that graph axes are aligned.

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

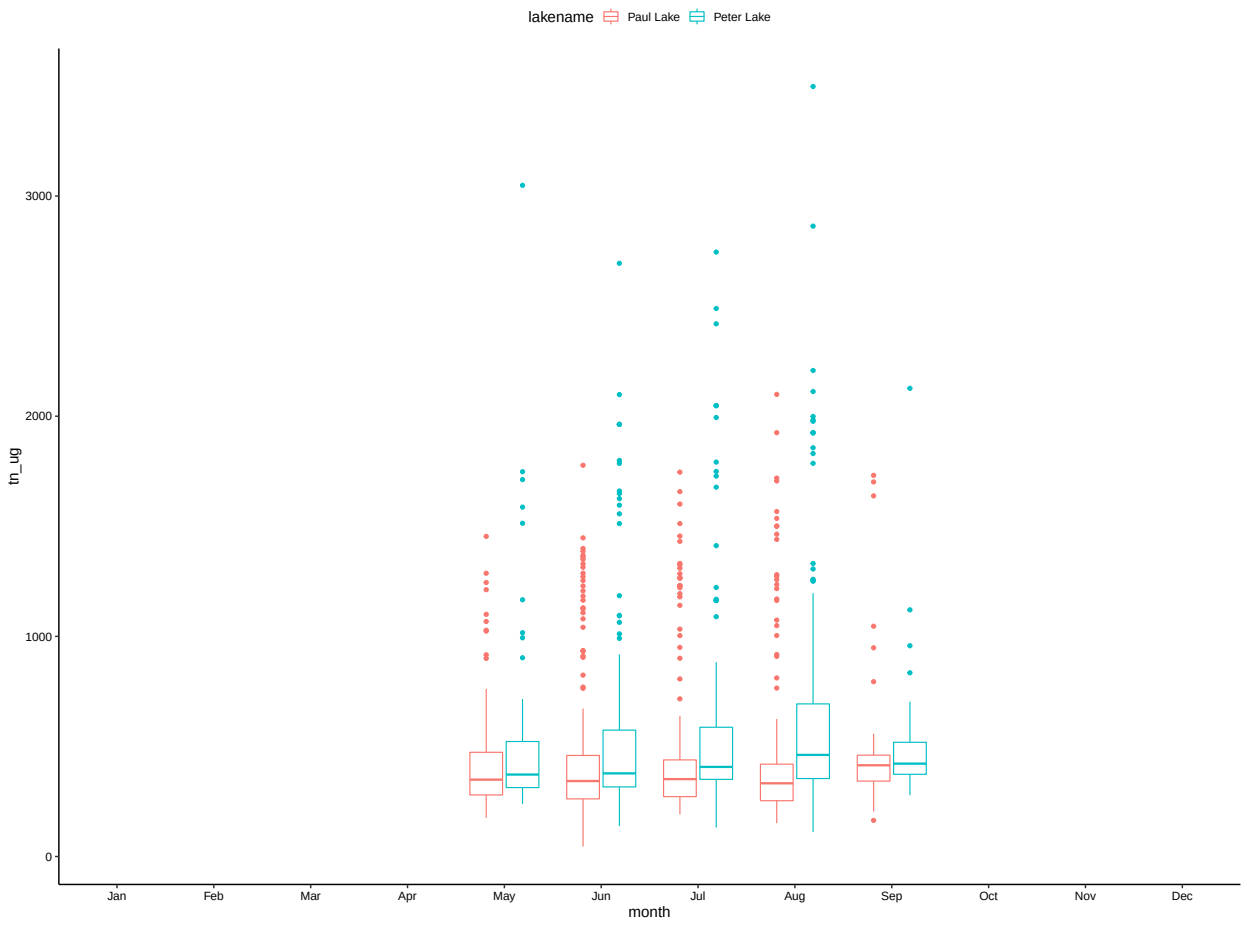
```
# 5
nutrient_data$month <- factor(nutrient_data$month, levels = c(1:12), labels = month.abb)
NutrientPlot_a <- ggplot(nutrient_data, aes(x = month, y = temperature_C)) + geom_boxplot(aes(color = la
    scale_x_discrete(drop = FALSE)
print(NutrientPlot_a)
```



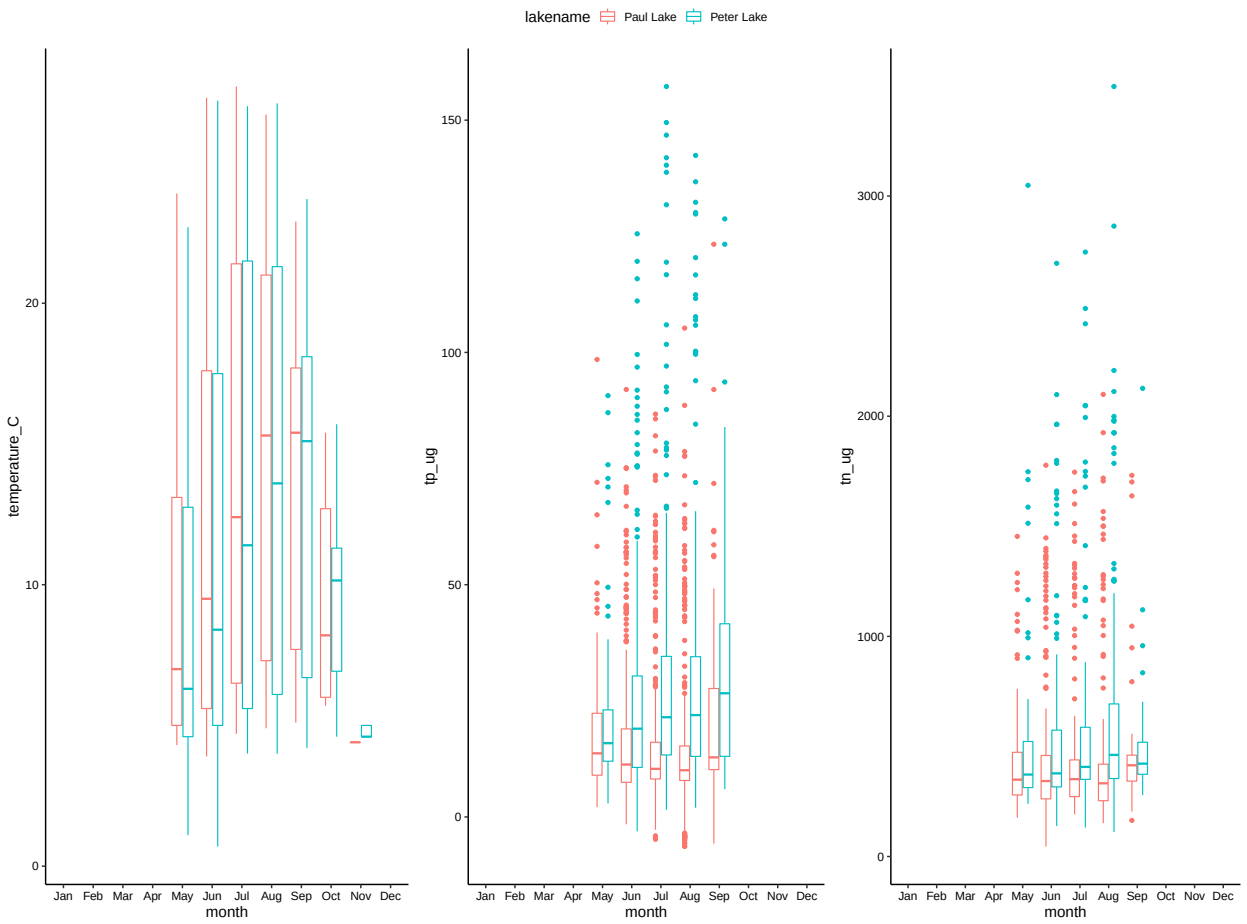

```
NutrientPlot_b <- ggplot(nutrient_data, aes(x = month, y = tp_ug)) + geom_boxplot(aes(color = lakename))
    scale_x_discrete(drop = FALSE)
print(NutrientPlot_b)
```




```
NutrientPlot_c <- ggplot(nutrient_data, aes(x = month, y = tn_ug)) + geom_boxplot(aes(color = lakename))
    scale_x_discrete(drop = FALSE)
print(NutrientPlot_c)
```




```
plot_grid(NutrientPlot_a + theme(legend.position = "none"), NutrientPlot_b, NutrientPlot_c +
    theme(legend.position = "none"), nrow = 1, align = "h", axis = "bt", rel_widths = c(1,
    1, 1))
```



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

Answer:

In terms of temperature, I could see in Paul Lake, from May to September, the median value continuously increased until reaching the peak in September, then decreased a lot in October and November. Also, in Peter Lake, I could see the same trend. When comparing temperatures between the two lakes, I could see that from May to September, the median temperature in Paul Lake is always larger than that in Peter Lake. While from October to November, the median temperature in Paul Lake is always smaller than that in Peter Lake.

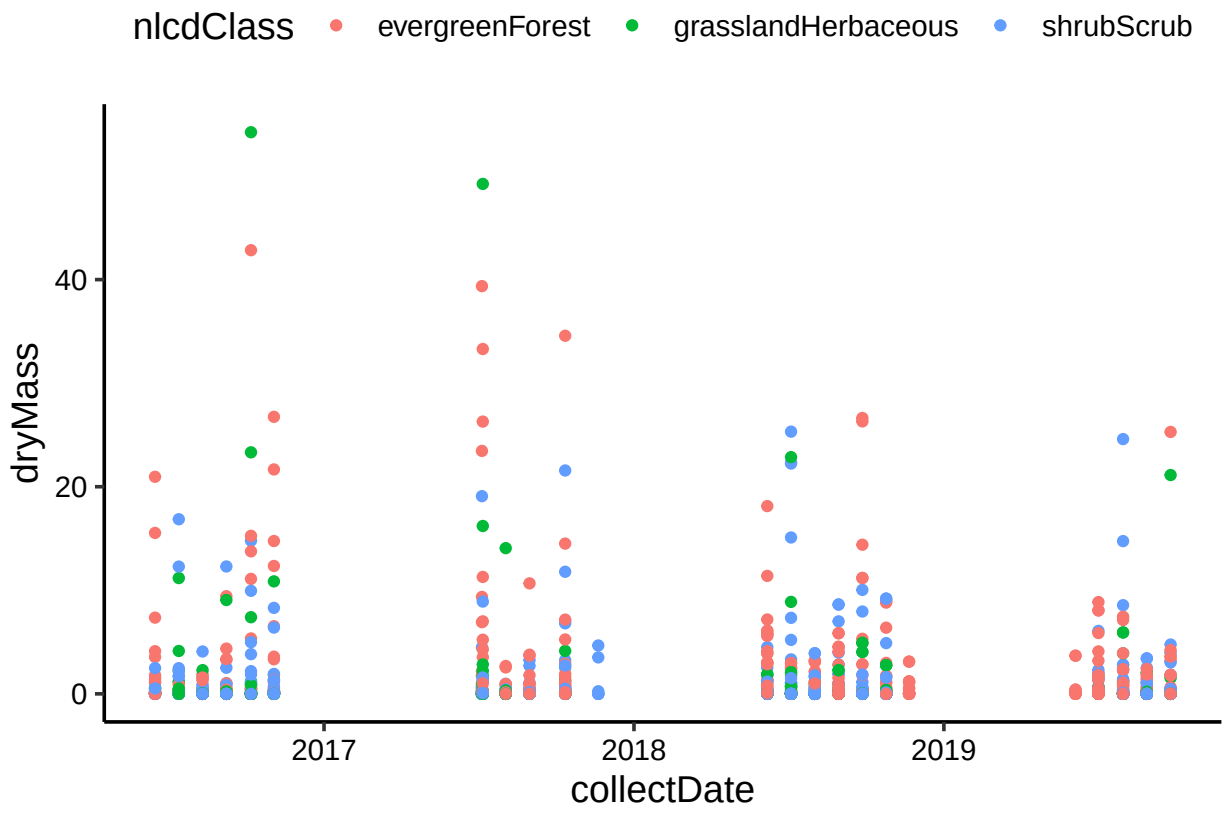
In terms of TP, I could see in Paul Lake, from May to September, the median TP value decreased to the bottom in August, then increased a little in September; while in Peter Lake, the median TP value experienced a continuous increase from May to September. When comparing TP values between the two lakes, I could see that the median, Q1 and Q3 TP values in Peter Lake are always larger than those in Paul Lake. In addition, I could see that the dispersion in Peter Lake's TP data is larger than that in Paul Lake's.

In terms of TN, I could see in Paul Lake, from May to July, the median TN value almost remained the same level then decreased a little in August, and increased to the peak in September; while in Peter Lake, the median TN value experienced a continuous increase from May to August then decreased a little in September. When comparing TN values between the two lakes, I could see that the median, Q1 and Q3 TN values in Peter Lake are always larger than those in Paul Lake. In addition, I could see that the dispersion in Peter Lake's TN data is larger than that in Paul Lake's.

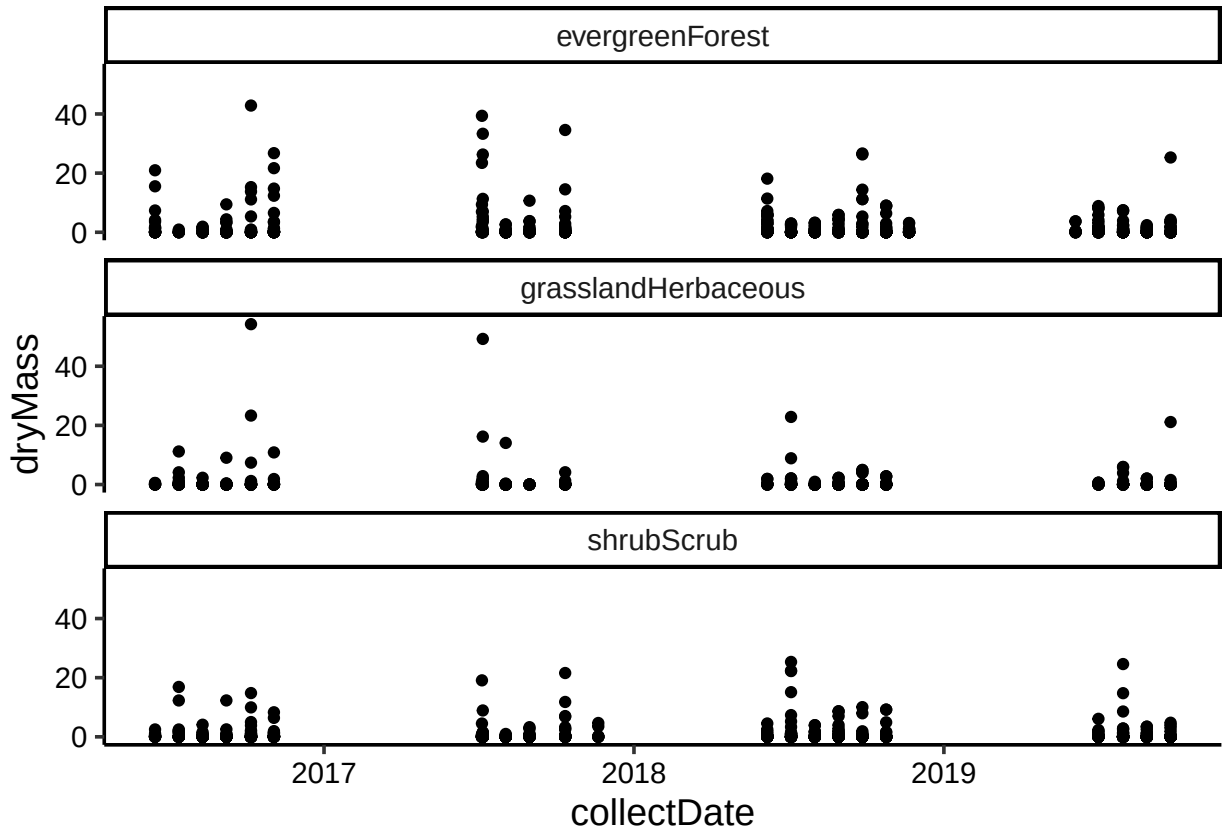
6. [Niwot Ridge] Plot a subset of the litter dataset by displaying only the "Needles" functional group. Plot the dry mass of needle litter by date and separate by NLCD class with a color aesthetic. (no need to adjust the name of each land use)

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

```
# 6
litterPlot.Ex6 <- ggplot(subset(litter_data, functionalGroup = "Needles"), aes(x = collectDate,
    y = dryMass)) + geom_point(aes(color = nlcdClass))
print(litterPlot.Ex6)
```




```
# 7
litterPlot.Ex7 <- ggplot(subset(litter_data, functionalGroup = "Needles"), aes(x = collectDate,
    y = dryMass)) + geom_point() + facet_wrap(vars(nlcdClass), nrow = 3)
print(litterPlot.Ex7)
```



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

Answer: Plot7 is more effective. Because it is easier for us to make contrast between three NLCD classes and to observe which class has more dry mass.