Assignment 5: Data Visualization

Li Jia Go

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 working directory
setwd("/home/guest/R/EDA-Fall2022/")

# load required packages
library(tidyverse)
library(lubridate)
library(cowplot)

# load and name PeterPaul
# dataset, convert strings to
# factor
PeterPaul <- read.csv("./Data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv",
    stringsAsFactors = TRUE)

# load and name Niwot Ridge
# litter dataset, convert
# strings to factor
NiwotRidge <- read.csv("./Data/Processed/NEON_NIWO_Litter_mass_trap_Processed.csv",
    stringsAsFactors = TRUE)
```

```r
# 2 read dates as date for
# both datasets
PeterPaul$sampledate <- as.Date(PeterPaul$sampledate)
PeterPaul$month <- as.factor(PeterPaul$month)

NiwotRidge$collectDate <- as.Date(NiwotRidge$collectDate)
```

Define your theme

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

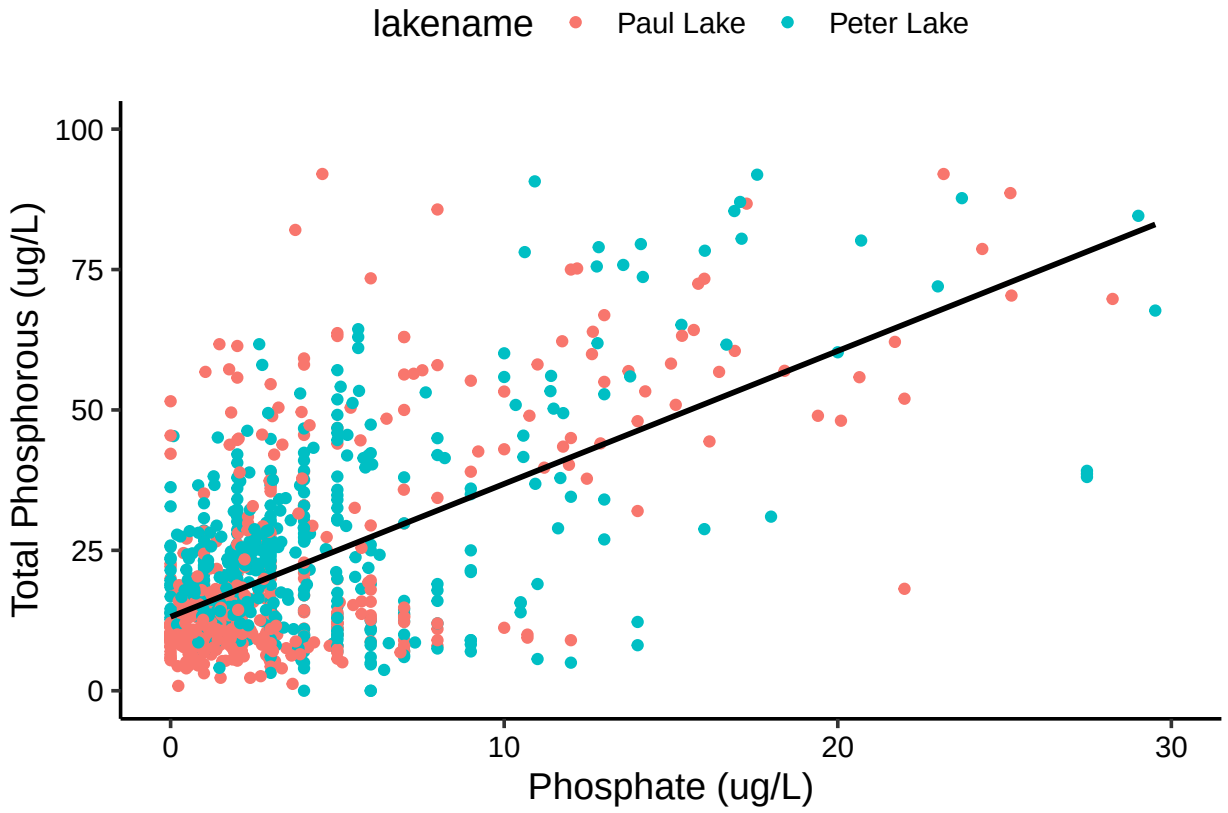
```r
# 3 building and setting a default theme
mytheme <- theme_classic(base_size = 14) + theme(axis.text = element_text(color = "black"),
    legend.position = "top")
theme_set(mytheme)
```

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()`).

```r
# 4 TP vs PO4 plot, with line of best fit in black
TPvPO4 <- ggplot(PeterPaul, aes(x = po4, y = tp_ug, color = lakename)) +
    geom_point() + geom_smooth(method = lm, color = "black",
    se = FALSE) + labs(x = "Phosphate (ug/L)", y = "Total Phosphorous (ug/L)") +
    xlim(0, 30) + ylim(0, 100)
print(TPvPO4)
```



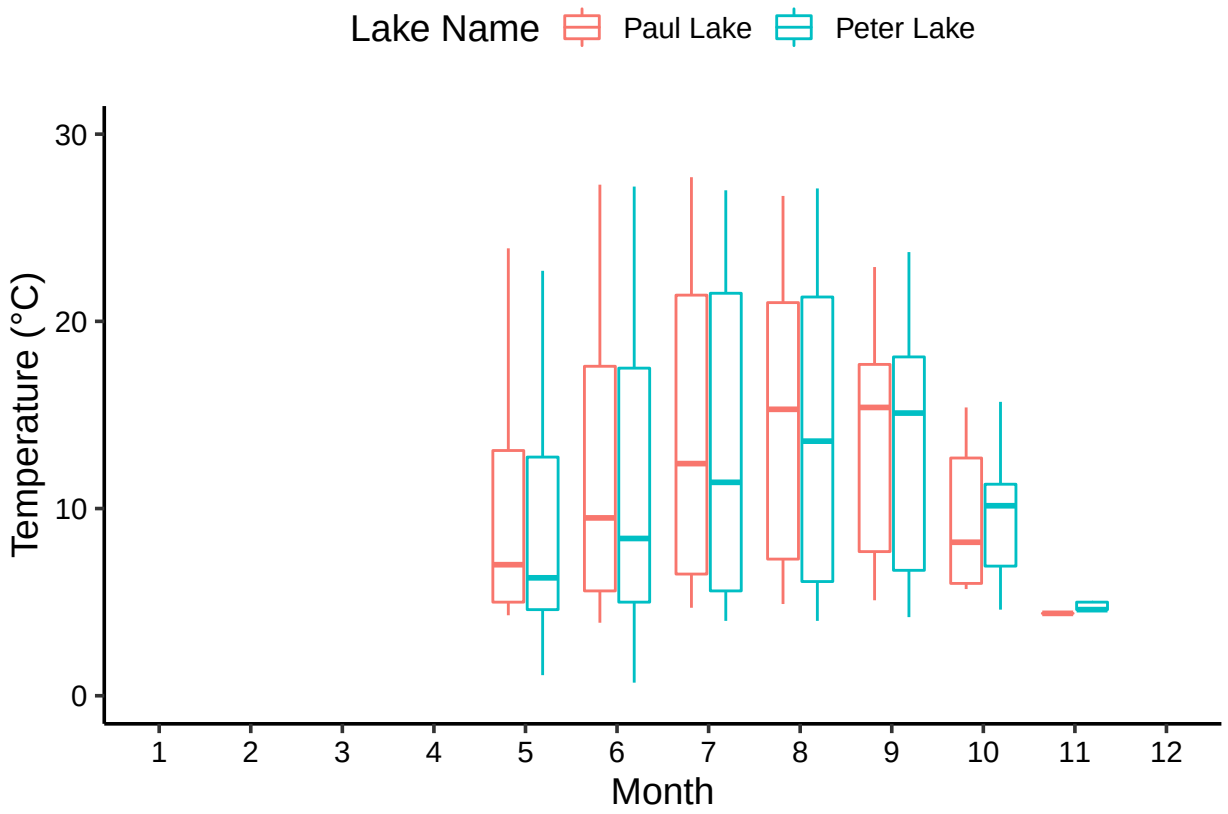
5. [NTL-LTER] Make three separate boxplots of (a) temperature, (b) TP, and

(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

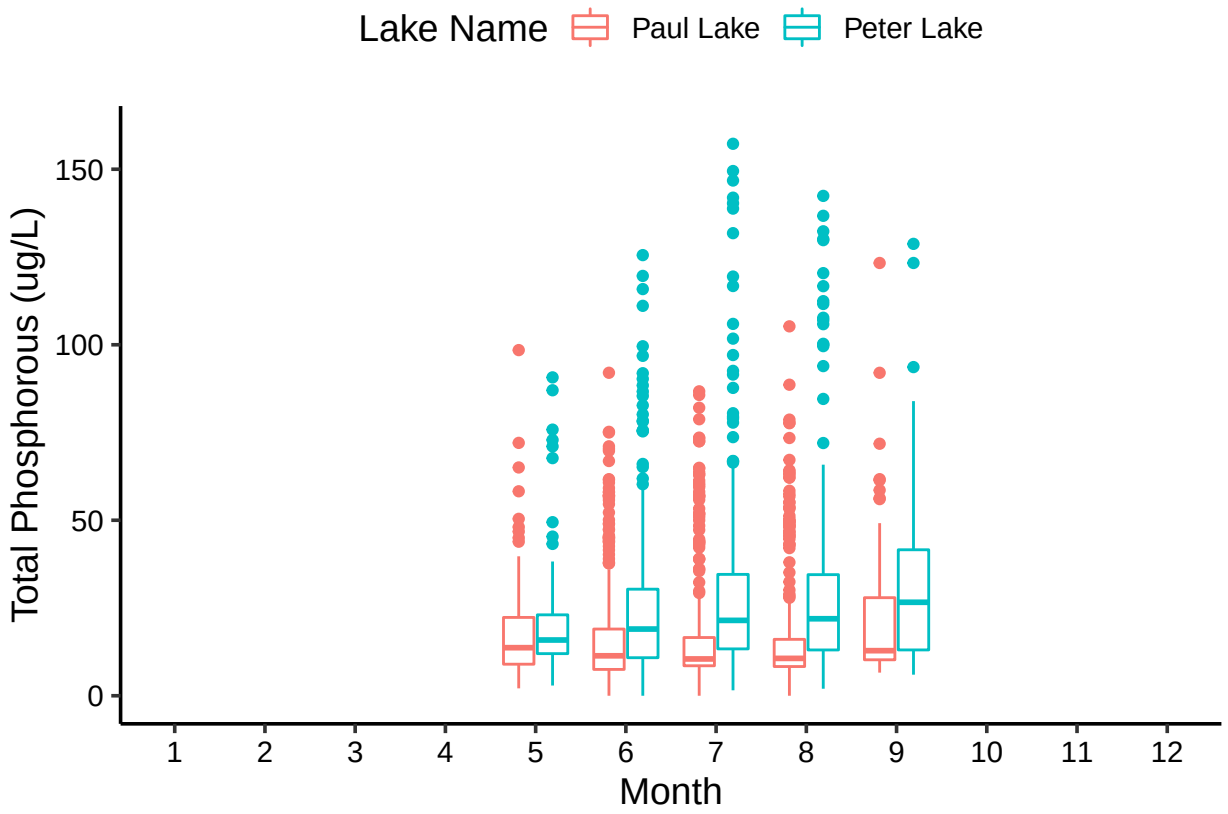
```
# 5(a) Temperature vs month boxplot. Reading months as
# factors. setting limits for y axis, and ensuring all
# months for x axis shows even without data in some months
Temp.monthplot <- ggplot(PeterPaul, aes(x = factor(month, levels = c(1:12)),
    y = temperature_C, color = lakename)) + geom_boxplot() +
    ylim(0, 30) + scale_x_discrete(drop = FALSE) + labs(x = "Month",
    y = "Temperature (°C)", color = "Lake Name")

print(Temp.monthplot)
```



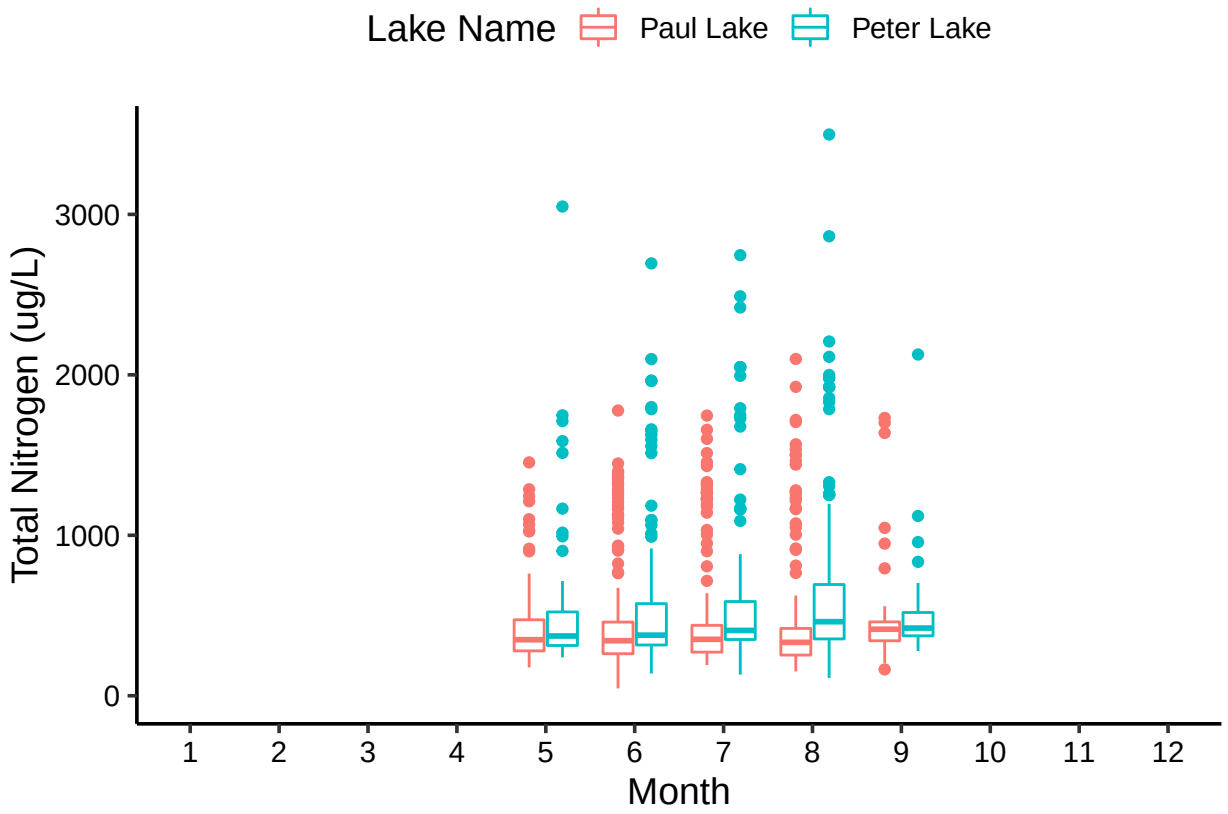

```
# 5(b) TP vs month boxplot
TP.monthplot <- ggplot(PeterPaul, aes(x = factor(month, levels = c(1:12)),
    y = tp_ug, color = lakename)) + geom_boxplot() + ylim(0,
    160) + scale_x_discrete(drop = FALSE) + labs(x = "Month",
    y = "Total Phosphorous (ug/L)", color = "Lake Name")

print(TP.monthplot)
```




```
# 5(c) TN vs month boxplot
TN.monthplot <- ggplot(PeterPaul, aes(x = factor(month, levels = c(1:12)),
    y = tn_ug, color = lakename)) + geom_boxplot() + ylim(0,
    3500) + scale_x_discrete(drop = FALSE) + labs(x = "Month",
    y = "Total Nitrogen (ug/L)", color = "Lake Name")

print(TN.monthplot)
```



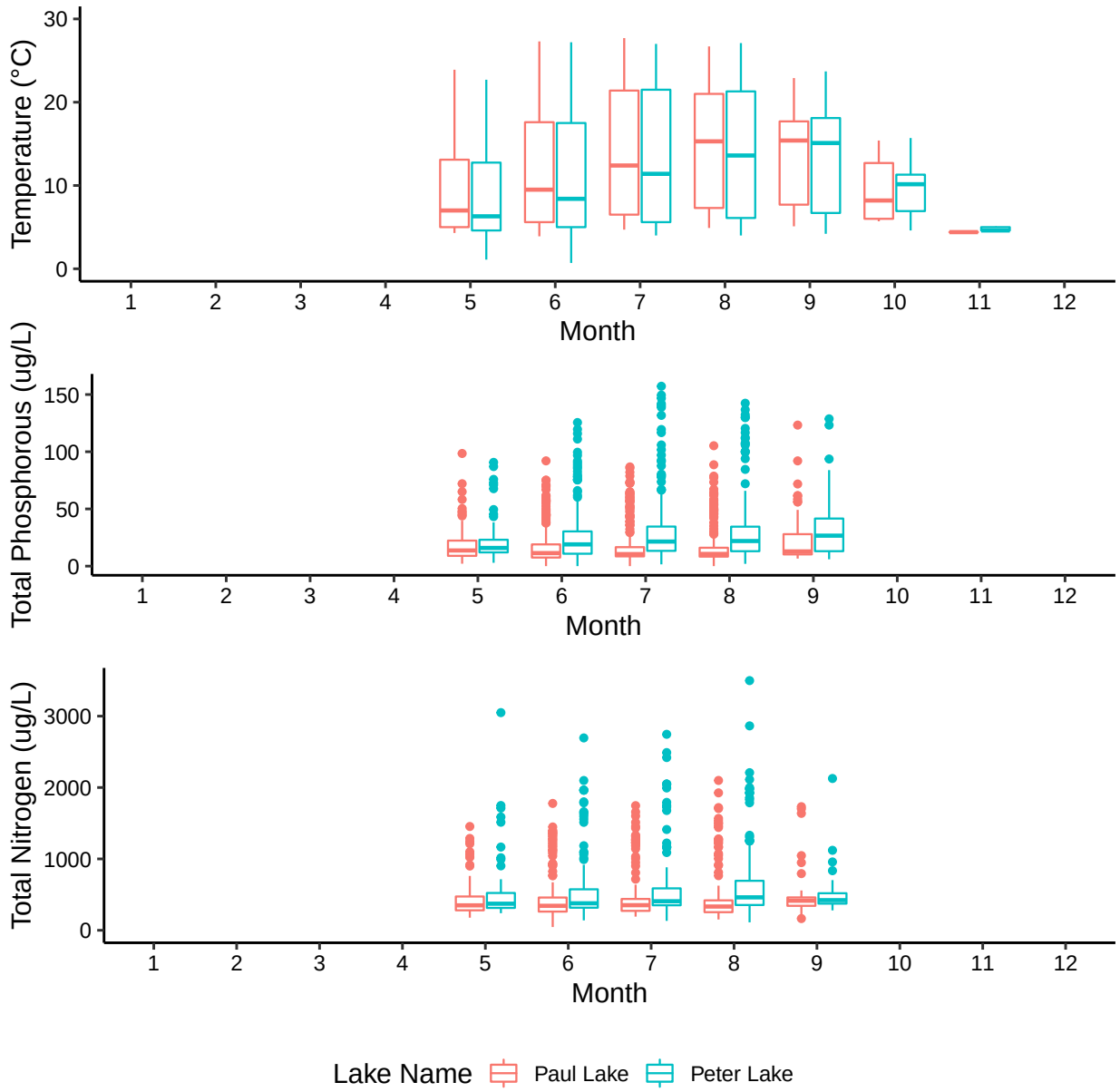

```
# 5(d) creating all 3 plots without legends
Temp.monthplot.NL <- Temp.monthplot + theme(legend.position = "none")

TP.monthplot.NL <- TP.monthplot + theme(legend.position = "none")

TN.monthplot.NL <- TN.monthplot + theme(legend.position = "none")

# combining all 3 plots with no legend and assigning them to a new object
combinedplots <- plot_grid(Temp.monthplot.NL, TP.monthplot.NL, TN.monthplot.NL, nrow = 3,
    align = "h", rel_heights = c(1.25, 1))

# extracting the legend from original temperature plot (with legend)
legend <- get_legend(Temp.monthplot + guides(color = guide_legend(nrow = 1)))
# cowplot combining all 3 plots with one single legend
plot_grid(combinedplots, legend, ncol = 1, rel_heights = c(1, 0.1))
```



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

Answer: Generally, we see wide temperature ranges for both lakes across all the months, though the median temperatures seem to increase from late Spring to Summer and decrease from summer to Spring. Of the 2 lakes, it seems that Peter Lake is more eutrophic than Paul Lake, having higher median TP and TN values as compared to Paul Lake for all the months where there are observations. Furthermore, the outliers for Peter Lake seem to hit much higher values as compared to Paul Lake for both nutrients. The median TN values for both lakes seem to remain relatively constant across Spring, Summer and early Fall, although we see a slight increase in the TN range and median value for Peter Lake in August. For TP, median values for Paul Lake remains relatively constant throughout the seasons, while median values for Peter Lake seem to rise from Spring to Summer and early Fall, with the highest values of TP recorded in the height of summer (July).

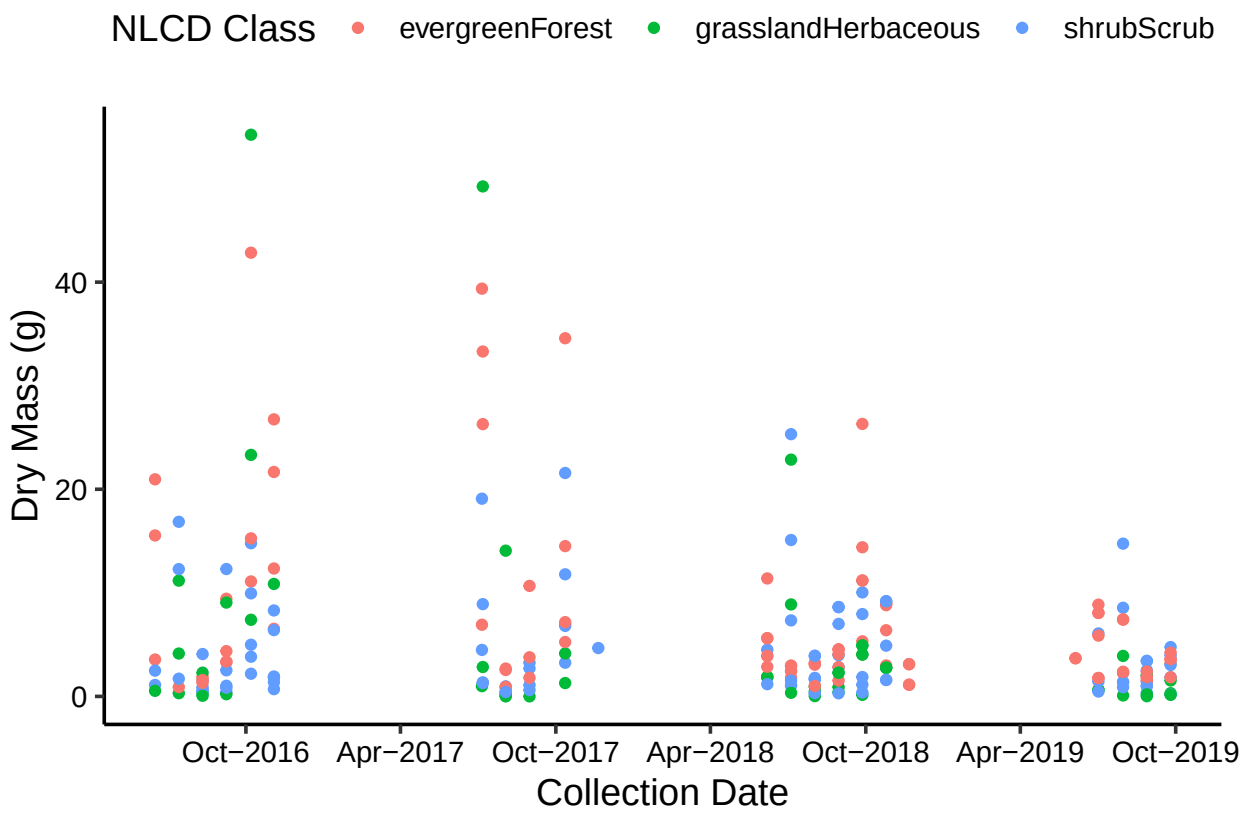
6. [Niwot Ridge] Plot a subset of the litter dataset by displaying only the "Needles" functional group. Plot

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.

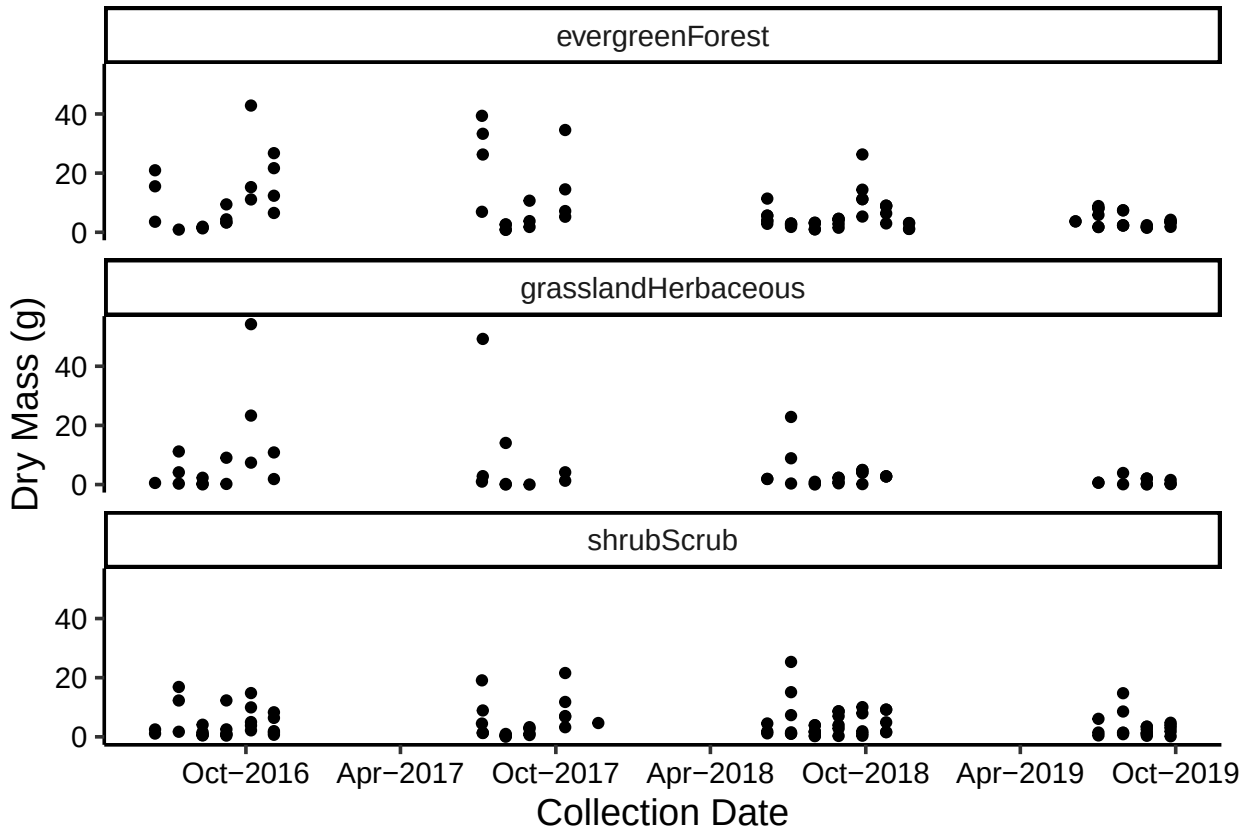
```
# 6 Drymass of needle dataset vs date, with NLCD classes
# separated by colour x axis ticks are in 6 month intervals
NeedleDryMass.Date <- ggplot(subset(NiwotRidge, functionalGroup ==
    "Needles"), aes(x = collectDate, y = dryMass, color = nlcdClass)) +
    geom_point() + scale_x_date(date_breaks = "6 months", date_labels = "%b-%Y") +
    labs(x = "Collection Date", y = "Dry Mass (g)", color = "NLCD Class")

print(NeedleDryMass.Date)
```




```
# 7 same plot separated by facets
NeedleDryMass.Date.Faceted <- ggplot(subset(NiwotRidge, functionalGroup ==
    "Needles"), aes(x = collectDate, y = dryMass)) + geom_point() +
    labs(x = "Collection Date", y = "Dry Mass (g)", color = "NLCD Class") +
    scale_x_date(date_breaks = "6 months", date_labels = "%b-%Y") +
    labs(x = "Collection Date", y = "Dry Mass (g)", color = "NLCD Class") +
    facet_wrap(vars(nlcdClass), nrow = 3)

print(NeedleDryMass.Date.Faceted)
```



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

Answer: 7 is more effective. We are better able to see the range of values for the dry mass of needle litter data for each NLCD site as compared to plot 6 where the points are clustered together on one axis.