Assignment 5: Data Visualization

Claire Mullaney

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Visualization

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Salk_A05_DataVisualization.Rmd") prior to submission.

The completed exercise is due on Tuesday, February 11 at 1:00 pm.

Set up your session

1. Set up your session. Verify your working directory and load the tidyverse and cowplot packages. Upload the NTL-LTER processed data files for nutrients and chemistry/physics for Peter and Paul Lakes (tidy and gathered) and the processed data file for the Niwot Ridge litter dataset.

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

```
#1
#Verifying directory, loading packages, and uploading files
getwd()
```

```
## [1] "/Users/clairemullaney/Desktop/ENV 872/Environmental_Data_Analytics_2020"
```

```
library(ggplot2)
library(viridis)
library(RColorBrewer)
library(tidyverse)
library(cowplot)


Peter_Paul_Nutrients <-
  read.csv("./Data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv")
Peter_Paul_Nutrients_Gathered <-
  read.csv("./Data/Processed/NTL-LTER_Lake_Nutrients_PeterPaulGathered_Processed.csv")
Niwot_Litter <- read.csv("./Data/Processed/NEON_NIWO_Litter_mass_trap_Processed.csv")

#2
#Looking at column names, checking the class of each date column, and
#converting those date columns to dates

###Peter_Paul_Nutrients
colnames(Peter_Paul_Nutrients)
```

```
##  [1] "lakename"        "year4"           "daynum"
##  [4] "month"           "sampledate"      "depth"
##  [7] "temperature_C"   "dissolvedOxygen" "irradianceWater"
```

```
## [10] "irradianceDeck"  "tn_ug"           "tp_ug"
## [13] "nh34"            "no23"            "po4"
```

```
class(Peter_Paul_Nutrients$sampledate)
```

```
## [1] "factor"
```

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

class(Peter_Paul_Nutrients$sampledate)
```

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

```
###Peter_Paul_Nutrients_Gathered
colnames(Peter_Paul_Nutrients_Gathered)
```

```
## [1] "lakename"      "year4"         "daynum"        "month"
## [5] "sampledate"    "depth"         "nutrient"      "concentration"
```

```
class(Peter_Paul_Nutrients_Gathered$sampledate)
```

```
## [1] "factor"
```

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

class(Peter_Paul_Nutrients_Gathered$sampledate)
```

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

```
###Niwot_Litter
colnames(Niwot_Litter)
```

```
##  [1] "plotID"           "trapID"           "collectDate"
##  [4] "functionalGroup"  "dryMass"          "qaDryMass"
##  [7] "subplotID"        "decimalLatitude"  "decimalLongitude"
## [10] "elevation"        "nlcdClass"        "plotType"
## [13] "geodeticDatum"
```

```
class(Niwot_Litter$collectDate)
```

```
## [1] "factor"
```

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

class(Niwot_Litter$collectDate)
```

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

Define your theme

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

```
#Defining a new theme
theme_5 <- theme_classic(base_size = 12) +
  theme(axis.text = element_text(color = "black"),
```

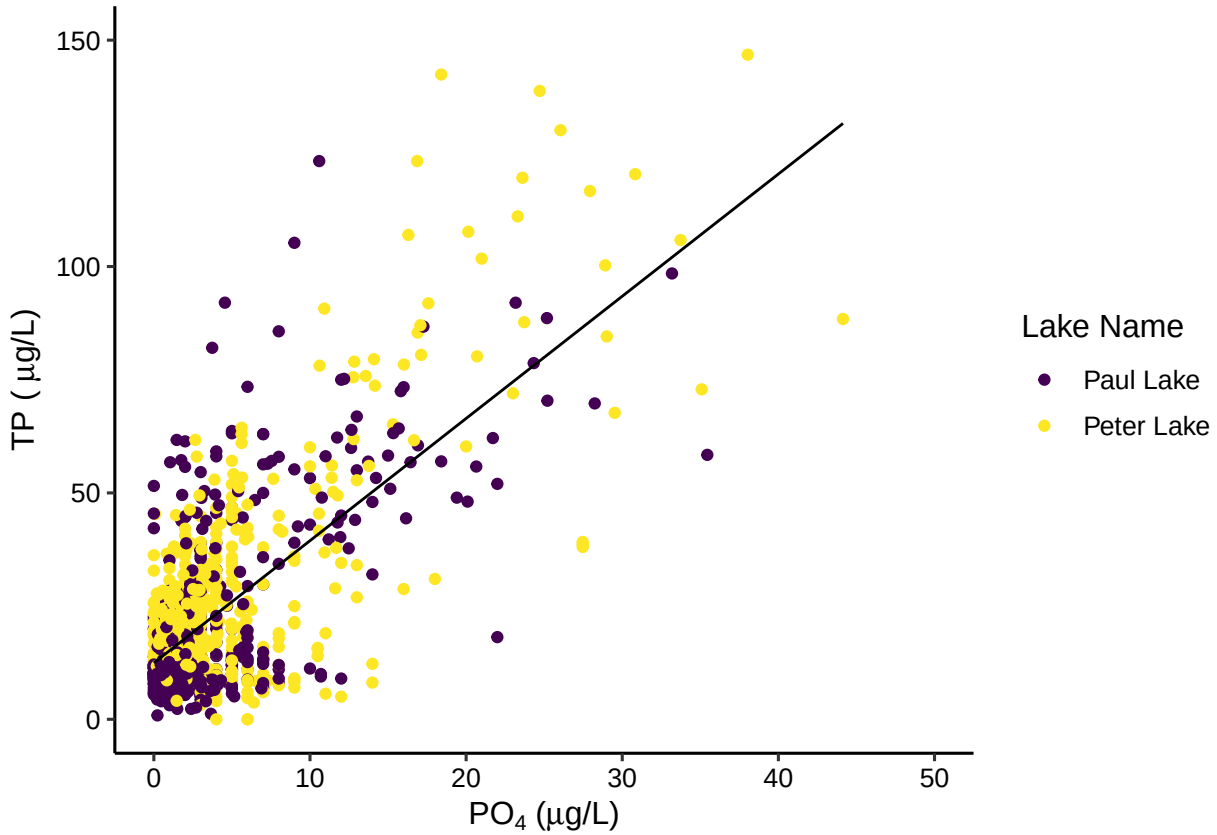
```
        legend.position = "right")

#Setting new theme as the default theme
theme_set(theme_5)
```

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 by phosphate, 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.

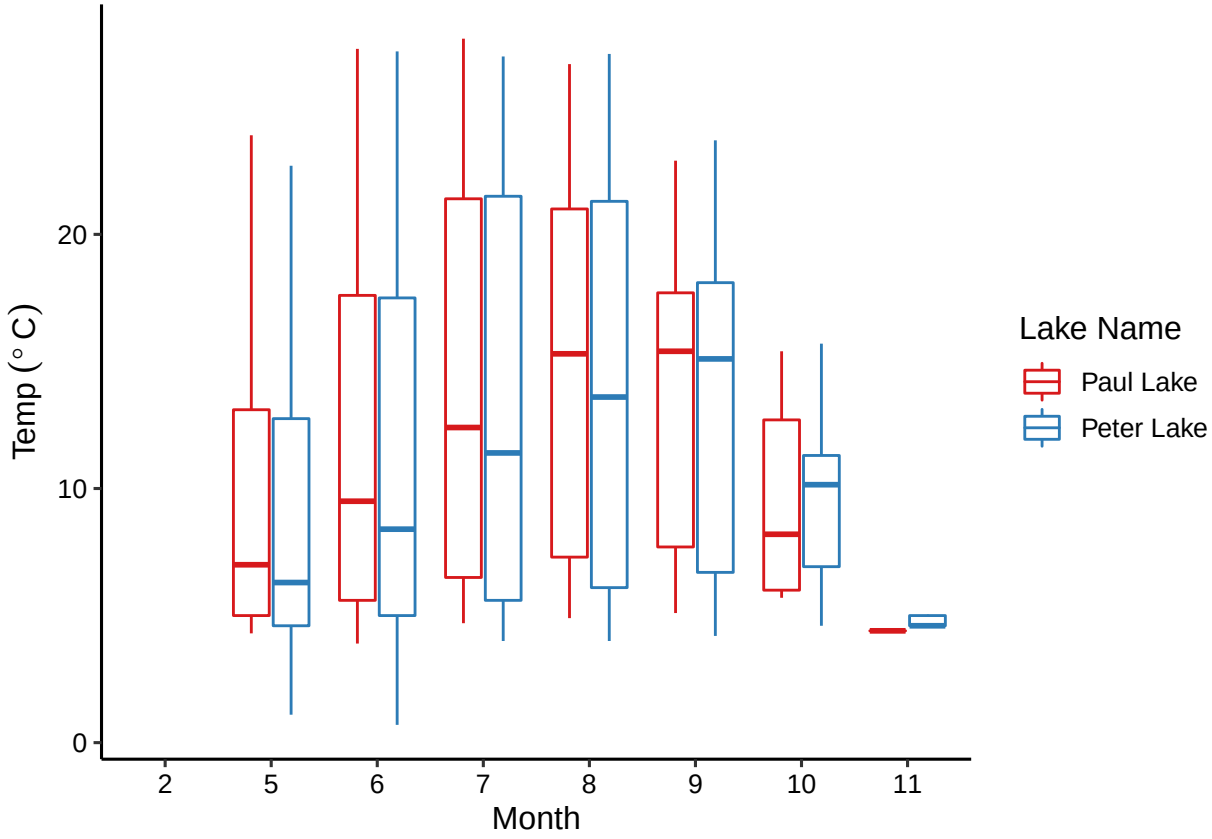
```
tp_vs_po4 <- ggplot(Peter_Paul_Nutrients, aes(y = tp_ug, x = po4, color = lakename)) +
  geom_point() +
  geom_smooth(method=lm, se = FALSE,
              color = "black", size = 0.5) +
  xlim(0, 50) +
  ylim(0, 150) +
  labs(y = expression(paste("TP ( ", mu, "g/L)")),
       x = expression(paste("PO"[4]* " (", mu, "g/L)")),
       color = "Lake Name") +
  scale_color_viridis(discrete = TRUE)

print(tp_vs_po4)
```



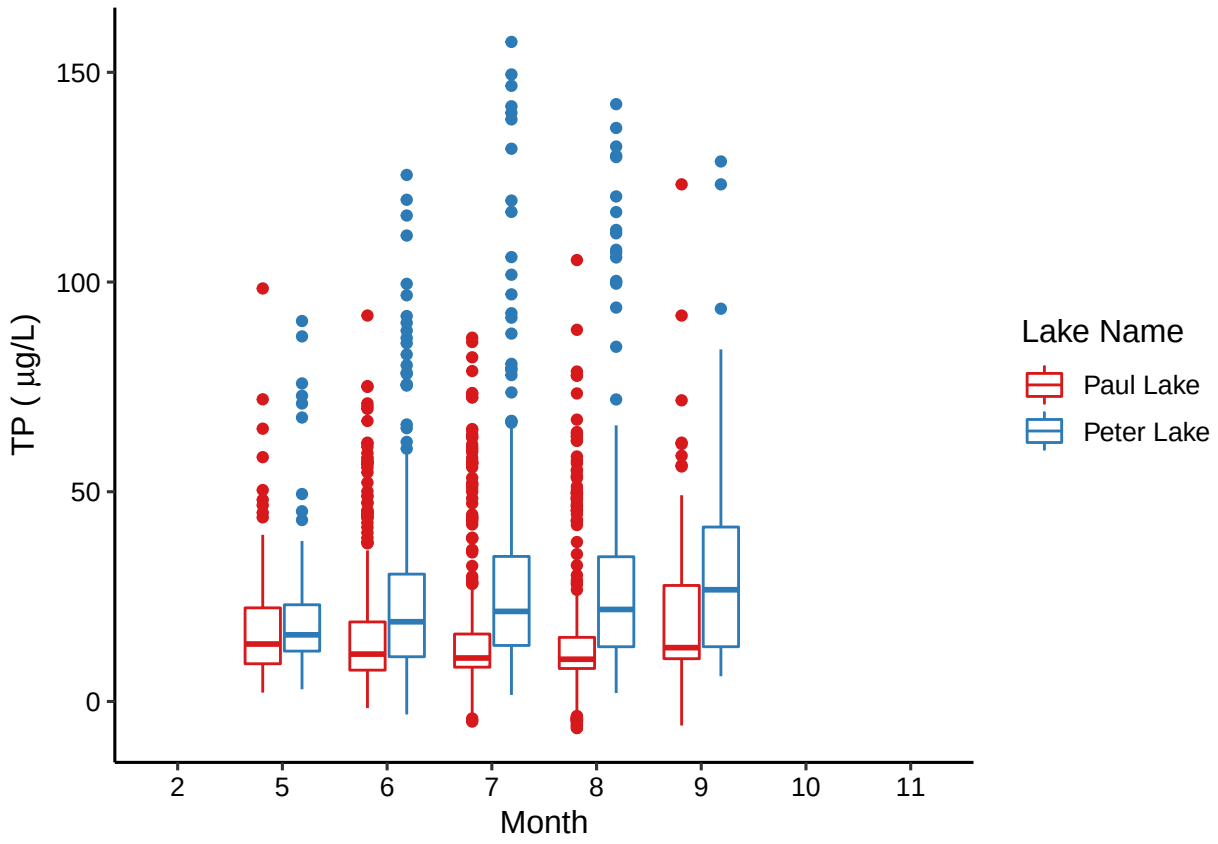
5. [NTL-LTER] Make three separate boxplots of (a) temperature, (b) TP, and (c) TN, with month as the x axis and lake as a color aesthetic. Then, create a cowplot that combines the three graphs. Make sure

that only one legend is present and that graph axes are aligned.

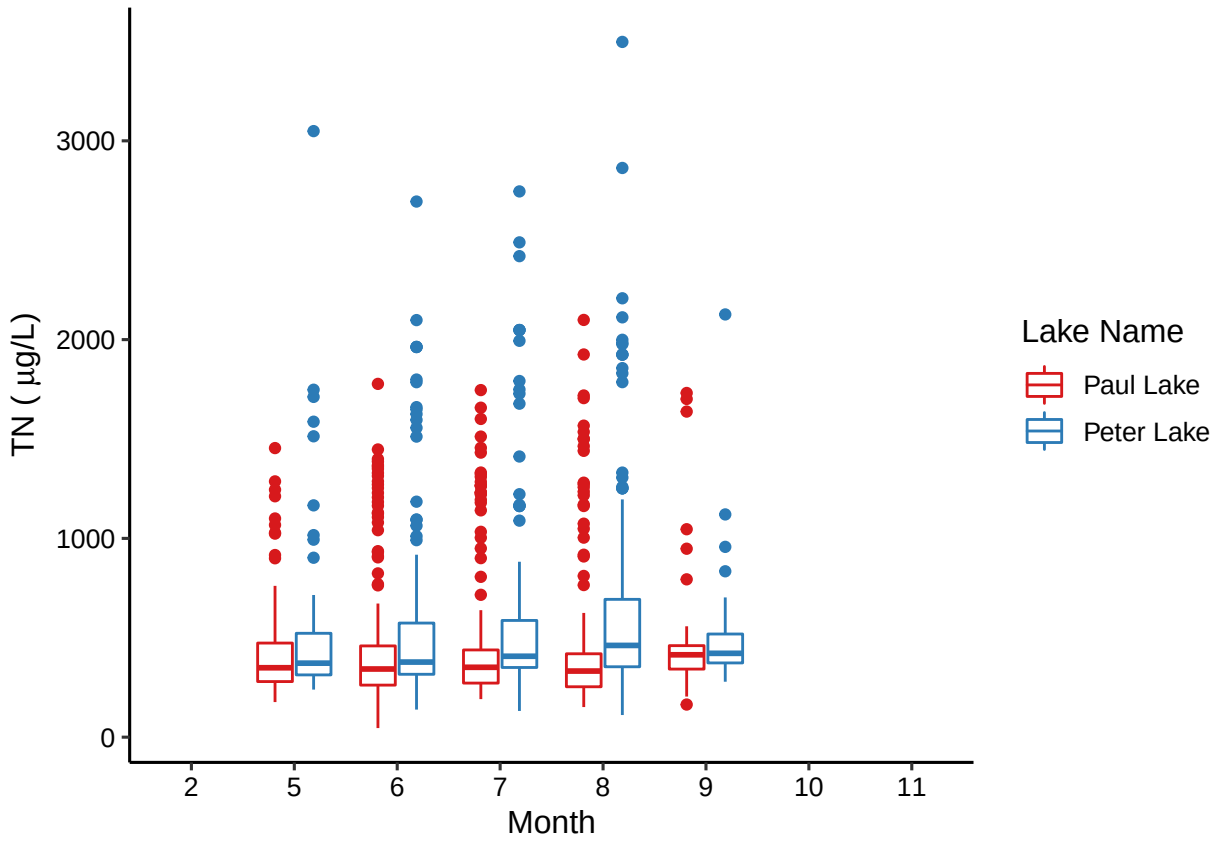
```
#Constructing boxplots

temp_box <- ggplot(Peter_Paul_Nutrients,
                   aes(y = temperature_C, x = as.factor(month),
                       color = lakename)) +
  geom_boxplot() +
  labs(y = expression("Temp " (degree~C)), x = "Month", color = "Lake Name") +
  scale_color_manual(values = c("#d7191c", "#2c7bb6"))

print(temp_box)
```



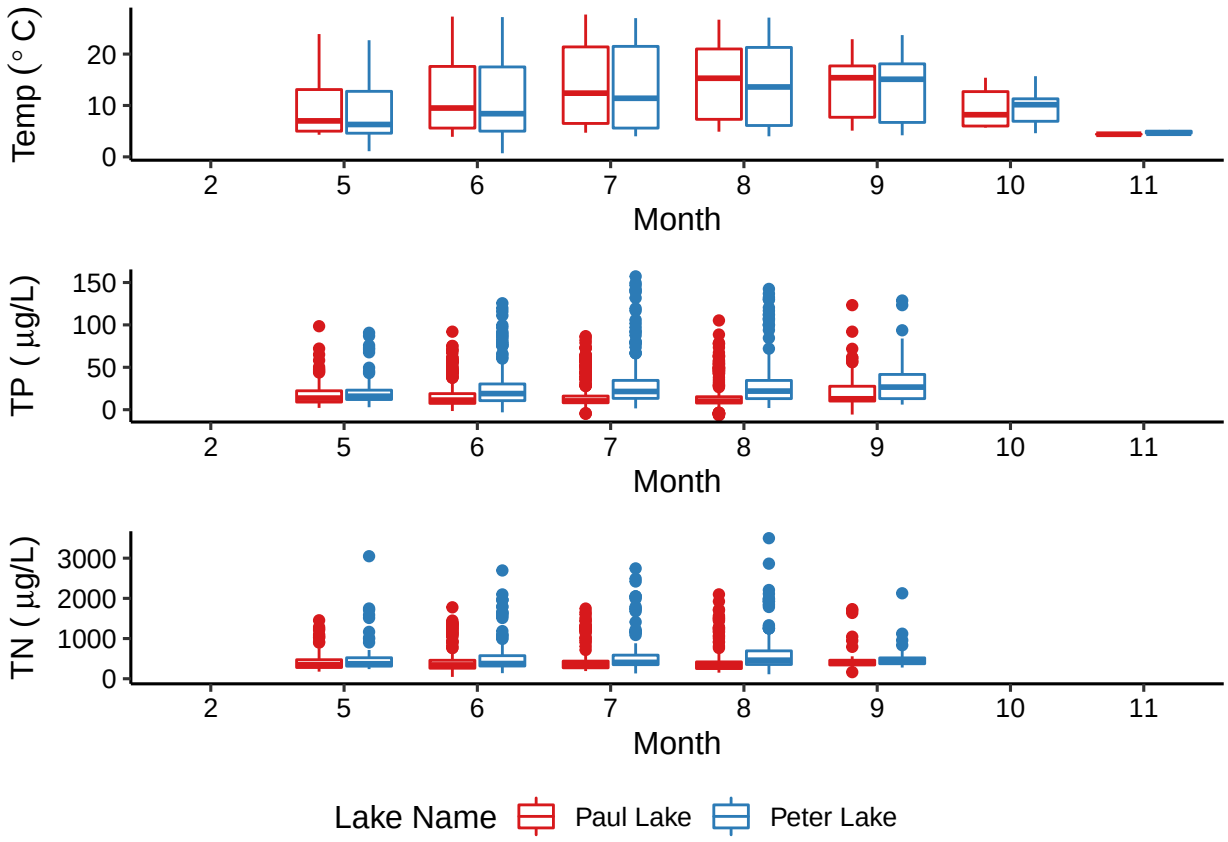

```
##########
tp_box <- ggplot(Peter_Paul_Nutrients,
                 aes(y = tp_ug, x = as.factor(month),
                     color = lakename)) +
        geom_boxplot() +
  labs(y = expression(paste("TP ( ", mu, "g/L)")),
       x = "Month", color = "Lake Name") +
  scale_color_manual(values = c("#d7191c", "#2c7bb6"))

print(tp_box)
```




```
##########
tn_box <- ggplot(Peter_Paul_Nutrients,
                 aes(y = tn_ug, x = as.factor(month),
                     color = lakename)) +
         geom_boxplot() +
    labs(y = expression(paste("TN ( ", mu, "g/L)")), x = "Month",
         color = "Lake Name") +
  scale_color_manual(values = c("#d7191c", "#2c7bb6"))

print(tn_box)
```




```
#Create a cowplot that combines the three graphs. Make sure that only
#one legend is present and that graph axes are aligned.

temp_tp_tn <- plot_grid(temp_box + theme(legend.position="none"),
                        tp_box + theme(legend.position="none"),
                        tn_box + theme(legend.position="none"),
                        nrow = 3, align = "hv")

legend <- get_legend(temp_box +
                       guides(color = guide_legend(nrow = 1)) +
                       theme(legend.position = "bottom"))

plot_grid(temp_tp_tn, legend, ncol = 1, rel_heights = c(1, .1))
```



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

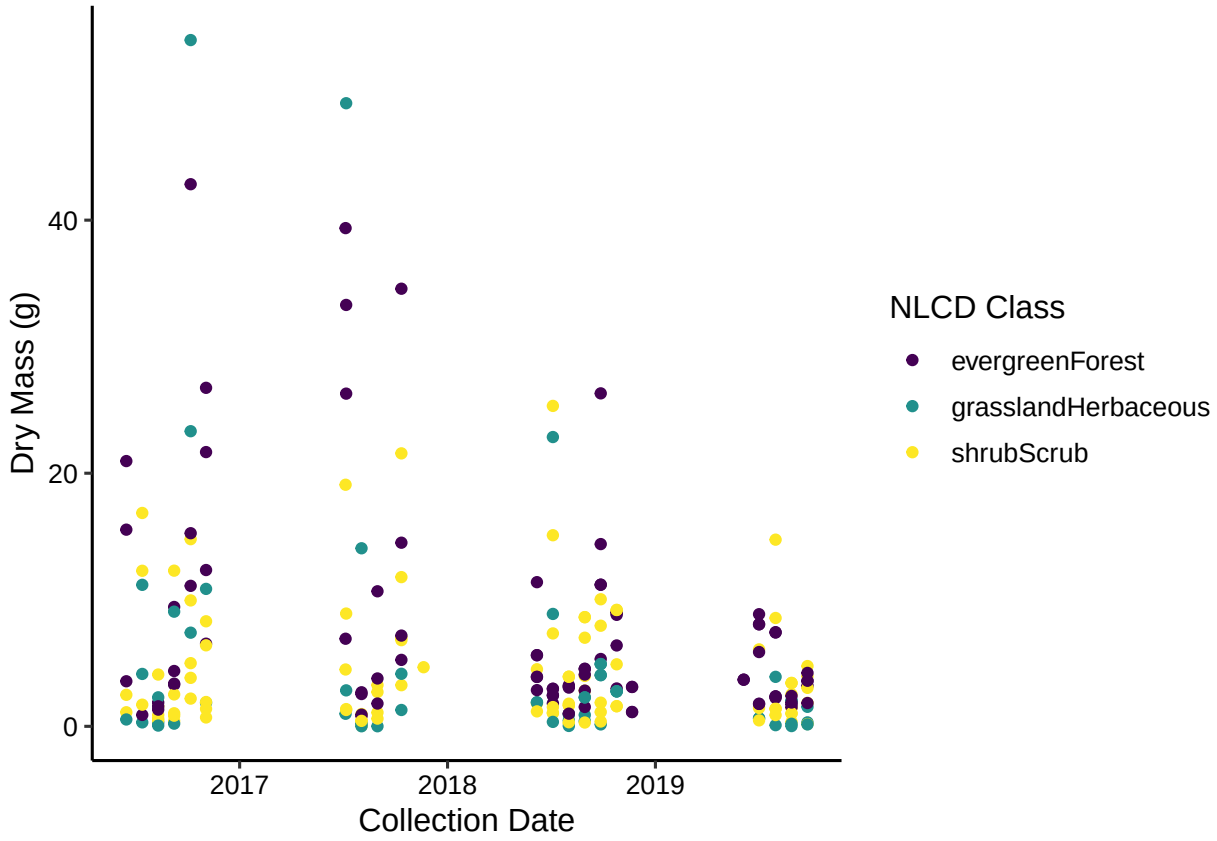
> Answer: Median temperature and temperature interquartile ranges (IQRs) increase from May through August and decrease from September through November. Peter Lake consistently has lower median temperatures, and often wider temperature IQRs, than Paul Lake (except in the month of October). The total amount of nitrogen in the lakes follows a similar pattern, with the median amount slightly increasing for both lakes from May through August and starting to decrease in September. Peter Lake often has larger median amounts of nitrogen, larger IQRs, and more positive skew than Paul Lake. Unlike temperature and total nitrogen, the median total amount of phosphorus in the lakes appears to increase from May all the way through September, the last month for which there is data. This increase is more prominent for Peter Lake, which also has higher median amounts of total phosphorus and larger IQRs. The distributions of total amounts of phosphorus for Paul Lake appear to be more positively skewed for any given month. For both Peter and Paul lake, there are many positive outliers in the distributions for both total nitrogen and total phosphorus in each month.

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)

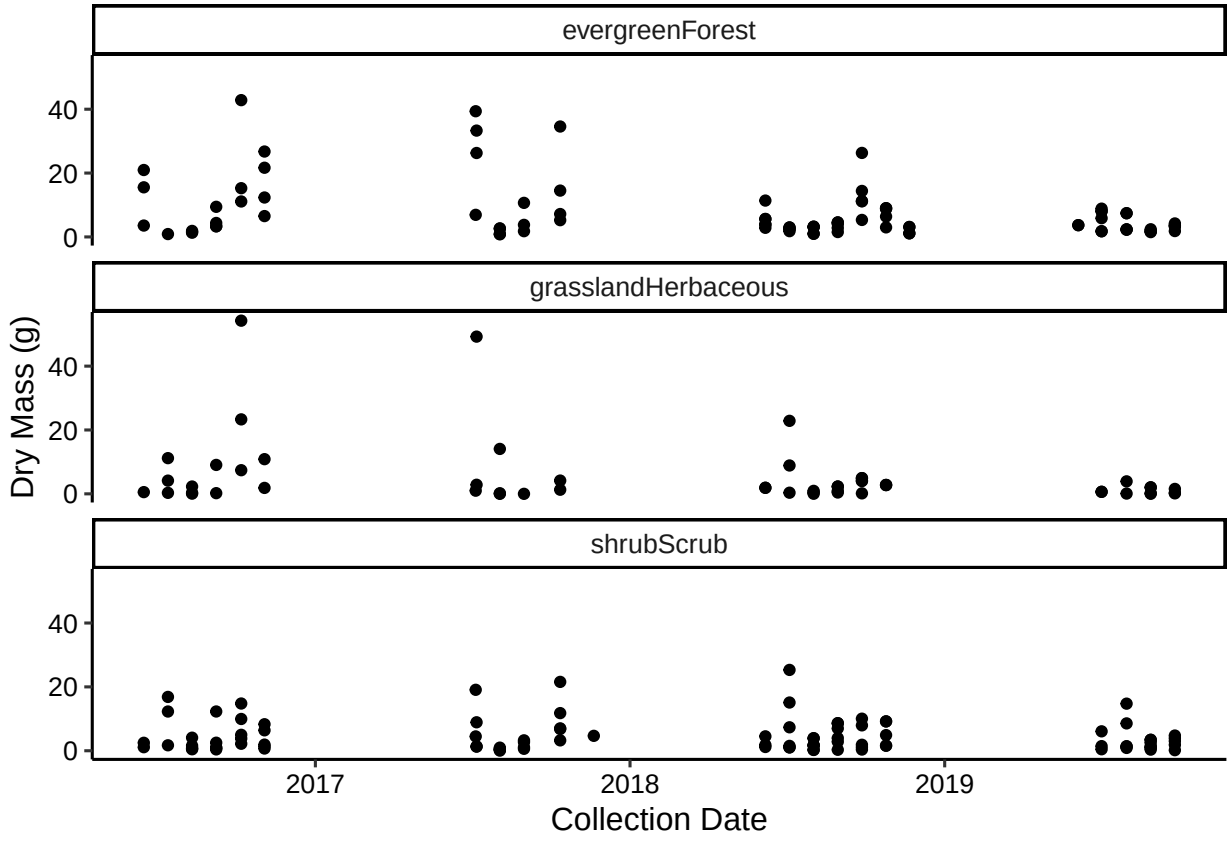
7. [Niwot Ridge] Now, plot the same plot but with NLCD classes separated into three facets rather than separated by color.

```
#6
#Color-coded graph
drymass_vs_date <- ggplot(subset(Niwot_Litter,
                                 functionalGroup == "Needles")) +
  aes(y = dryMass, x = collectDate, color = nlcdClass) +
  geom_point() +
```

```
  labs(y = "Dry Mass (g)", x = "Collection Date", color = "NLCD Class") +
   scale_color_viridis(discrete = TRUE)

print(drymass_vs_date)
```




```
#7
#Faceted graph

drymass_vs_date_fac <- ggplot(subset(Niwot_Litter,
                                     functionalGroup ==
                                       "Needles")) +
  aes(y = dryMass, x = collectDate) +
  geom_point() +
  facet_wrap(vars(nlcdClass), nrow = 3) +
  labs(y = "Dry Mass (g)", x = "Collection Date")

print(drymass_vs_date_fac)
```



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

Answer: I think plot 7 is more effective; when facet_wrap is used to create three individual graphs (one for each NLCD class), each in its own row, it is easy to see the dry masses of each NLCD type within any given year. These dry masses can be efficiently compared both across classes (by looking at a vertical segment of all three plots) and across years (by looking at each individual graph). In plot 6, even though the NLCD classes are separated by color, it is harder to collectively see the dry masses of each individual NLCD class for any given date range. While this difficulty does not result in comparisons across classes within a given date range being terribly cumbersome (although overlap in some data points does make these comparisons a bit more challenging than with plot 7), it does make examining changes in dry mass for one individual class across all years much more difficult in plot 6 than plot 7.