Assignment 5: Data Visualization

Anne Harshbarger

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., "Fay_A05_DataVisualization.Rmd") prior to submission.

The completed exercise is due on Tuesday, February 23 at 11:59 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 (both the tidy [`NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv`] and the gathered [`NTL-LTER_Lake_Nutrients_PeterPaulGathered_Processed.csv`] versions) 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
setwd("C:/Users/Anne/Documents/ENV872/Environmental_Data_Analytics_2021/Assignments")
getwd()
```

```
## [1] "C:/Users/Anne/Documents/ENV872/Environmental_Data_Analytics_2021/Assignments"
```

```
#install.packages("tidyverse", "cowplot", "ggplot2")
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts -------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(cowplot)
library(ggplot2)

#Create tidy lake nutrient dataset from csv
PeterPaul_Tidy <-
  read.csv(file = "../Data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv",
           stringsAsFactors = TRUE)

#create gathered lake nutrient dataset from csv
PeterPaul_Gathered <-
  read.csv(file = "../Data/Processed/NTL-LTER_Lake_Nutrients_PeterPaulGathered_Processed.csv",
           stringsAsFactors = TRUE)

#create Niwot Ridge dataset from CSV
Niwot <-
  read.csv(file = "../Data/Processed/NEON_NIWO_Litter_mass_trap_Processed.csv",
           stringsAsFactors = TRUE)

#2
class(PeterPaul_Tidy$sampledate) #check class of data column
```

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

```
PeterPaul_Tidy$sampledate <- as.Date(PeterPaul_Tidy$sampledate,
                                     format = "%Y-%m-%d") #set column as date

class(PeterPaul_Gathered$sampledate) #check class of date column
```

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

```
PeterPaul_Gathered$sampledate <- as.Date(PeterPaul_Gathered$sampledate,
                                     format = "%Y-%m-%d") #set column as date

class(Niwot$collectDate) #check class of date column
```

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

```
Niwot$collectDate <- as.Date(Niwot$collectDate,
                             format = "%Y-%m-%d") #set column as date
```

Define your theme

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

```
mytheme <- theme_bw() + theme( #create theme
  text = element_text(color = "black", size = 12), #set text size and color
  axis.text = element_text(color = "black", size = 12), #set axis text
  legend.position = "bottom", #put legend at bottom
  plot.background = element_rect("white", linetype = NULL), #create white background for whole plot
  panel.background = element_rect("white", linetype = NULL), # create white background for chart panel
  axis.line = element_line(color = "midnightblue", size = 1.5)) #set axis color and size

theme_set(mytheme) #set as default theme
```

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.

```
#create scatterplot with line of best fit for total p and po4
totalp_byphosphate <- ggplot(data = PeterPaul_Tidy,
                             aes(x = po4, y = tp_ug, color = lakename)) +
  geom_point(size = 1) +
  geom_smooth(method = lm, color = "black") +
  xlim(0,45) +
  labs(x = "Phosphate", y = "Total phosphorus (ug)", title = "Phosphorus by phosphate in Peter and Paul
print(totalp_byphosphate)
```

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

## Warning: Removed 21947 rows containing non-finite values (stat_smooth).

## Warning: Removed 21947 rows containing missing values (geom_point).
```

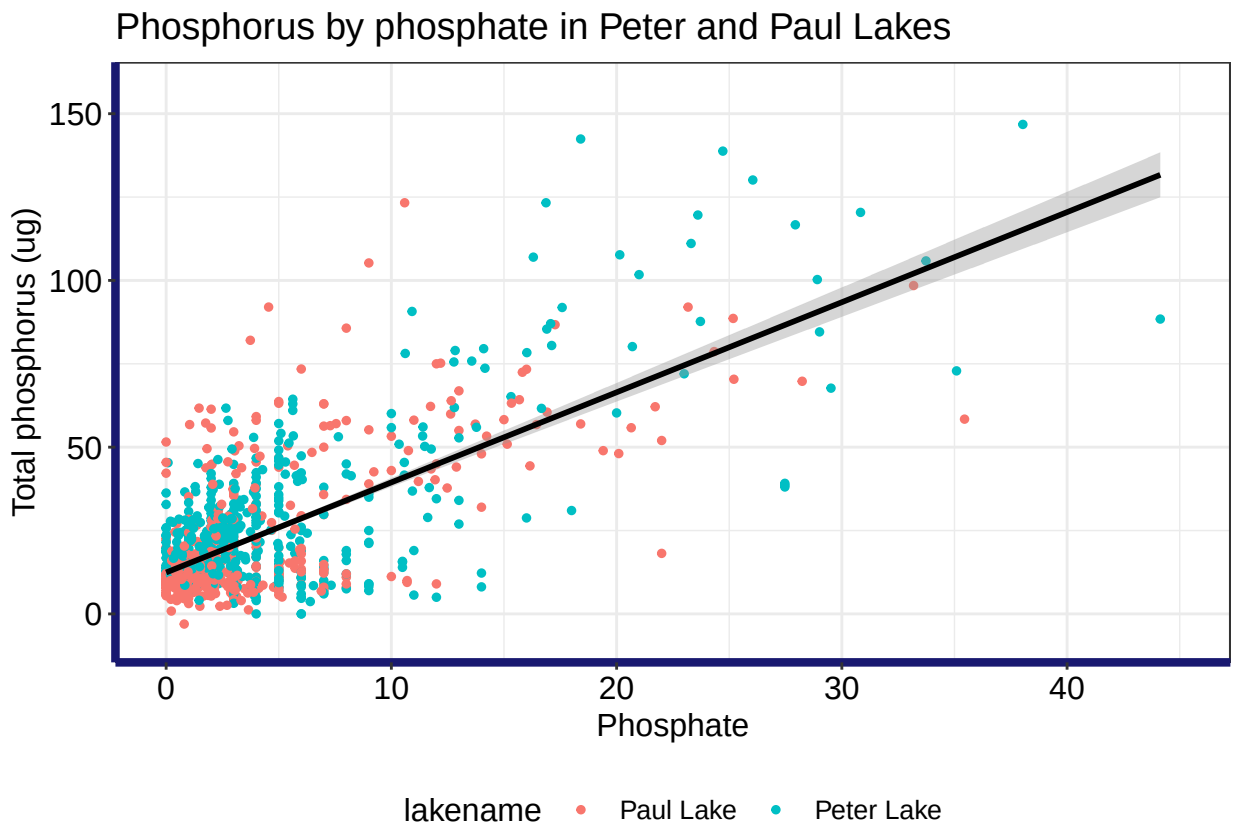

Figure 1: Relationship between phosphate and phosphorus in Peter Lake (blue) and Paul Lake (pink)

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.

```
#Convert month to factor
PeterPaul_Tidy$month <- as.factor(PeterPaul_Tidy$month)

#Create boxplots of temperature
box_temp <- ggplot(data = PeterPaul_Tidy,
                   aes(x = month, y = temperature_C, color = lakename)) +
  geom_boxplot() +
  labs(x = "Month", y = "Temperature (degrees C)",
       title = "Monthly temperatures in Peter and Paul Lakes")
print(box_temp)
```

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

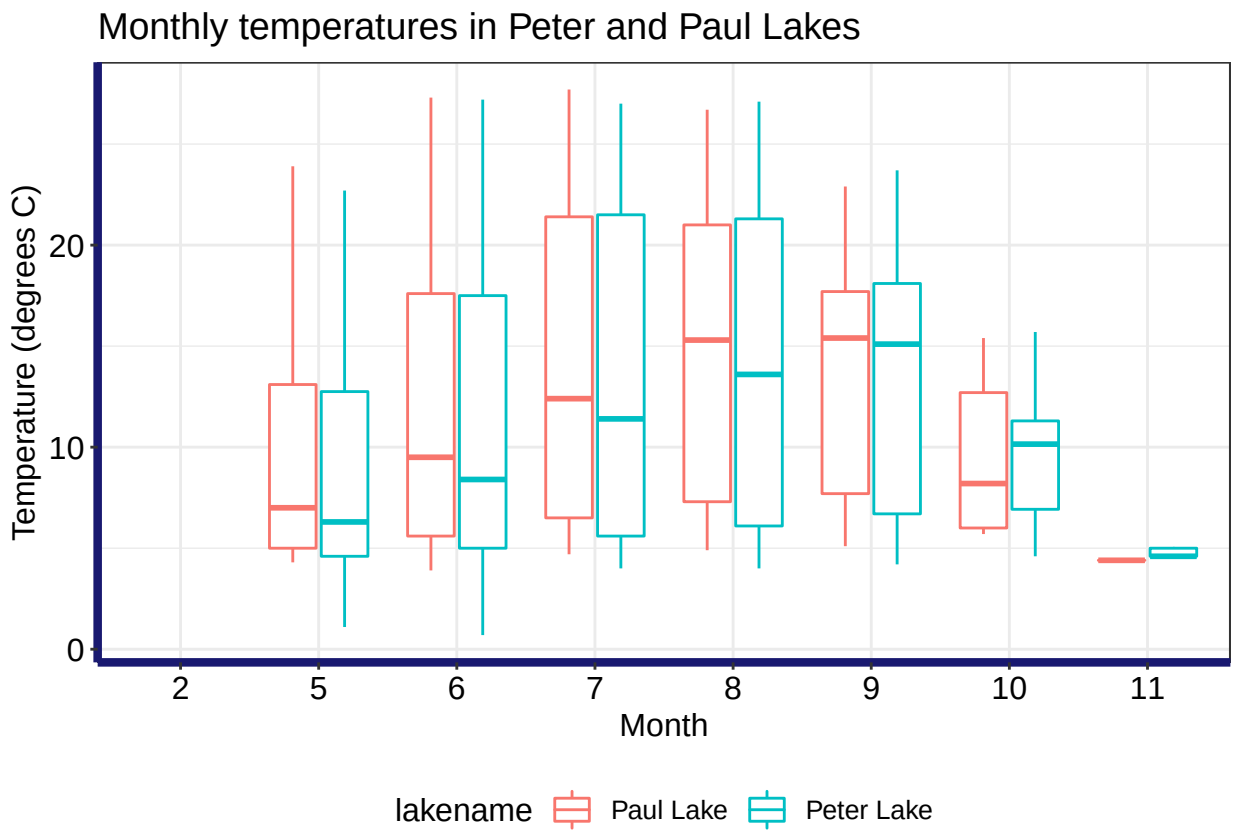

Figure 2: Temperature by month for Peter Lake (blue) and Paul Lake (pink)

```
#create boxplots of total phosphorus
box_TP <- ggplot(data = PeterPaul_Tidy,
                 aes(x = month, y = tp_ug, color = lakename)) +
  geom_boxplot() +
  labs(x = "Month", y = "Total Phosphorus (ug)",
       title = "Total phosphorus by month in Peter and Paul Lakes")
print(box_TP)
```

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

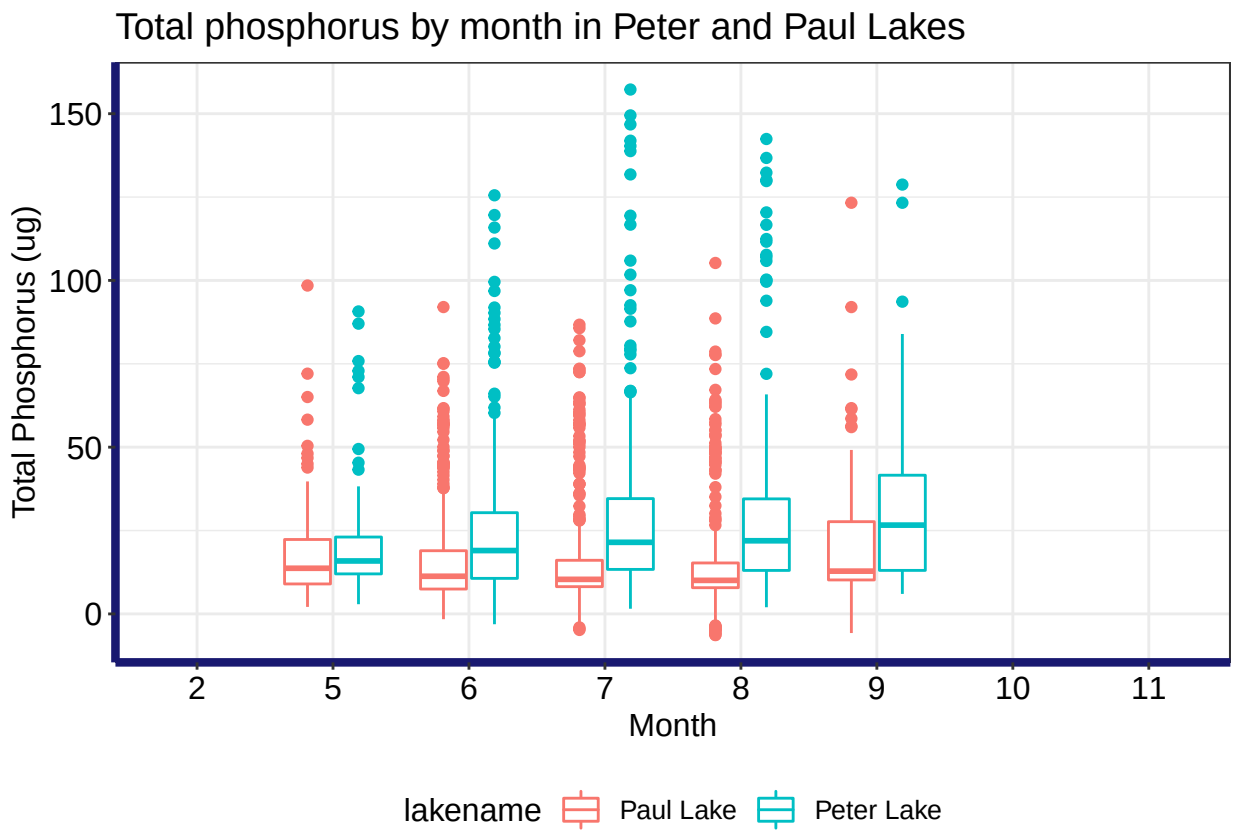

Figure 3: Total nitrogen (ug) by month for Peter Lake (blue) and Paul Lake (pink)

```
#create boxplots of total nitrogen
box_TN <- ggplot(data = PeterPaul_Tidy,
                 aes(x =month, y = tn_ug, color = lakename)) +
  geom_boxplot() +
  labs(x = "Month", y = "Total Nitrogen (ug)",
       title = "Total nitrogen by month in Peter and Paul Lakes")
print(box_TP)
```

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

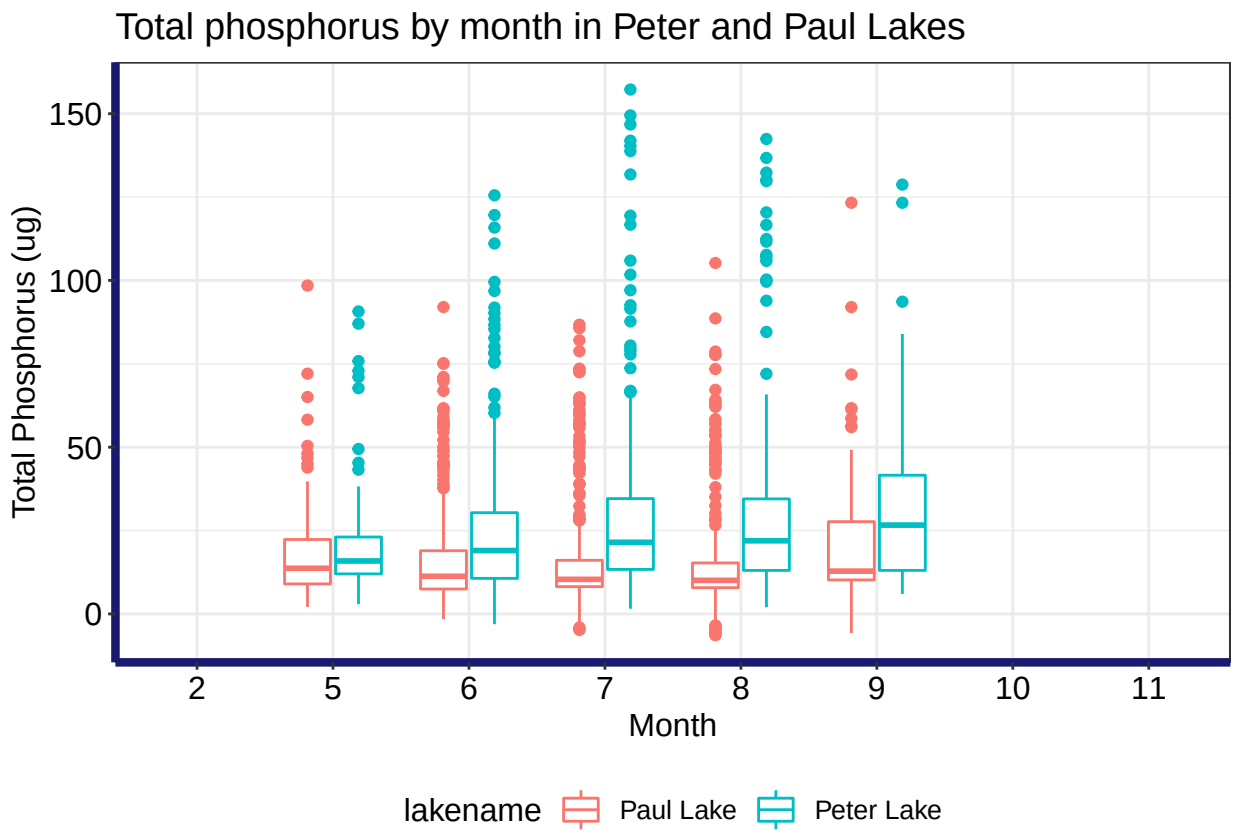

Figure 4: Total phosphorus (ug) by month for Peter Lake (blue) and Paul Lake (pink)

```
#duplicate 3 plots, removing 2 legends and titles to combine
box_temp_clean <- ggplot(data = PeterPaul_Tidy,
                  aes(x = month, y = temperature_C, color = lakename)) +
  geom_boxplot() +
  theme(legend.position="none") +
  labs(x = "Month", y = "Temperature (*C)",
       title = "Monthly Temperature, N, and P in Peter and Paul Lakes")

box_TP_clean <- ggplot(data = PeterPaul_Tidy,
                  aes(x = month, y = tp_ug, color = lakename)) +
  geom_boxplot() +
  theme(legend.position="none") +
  labs(x = "Month", y = "Total P (ug)")

box_TN_clean <- ggplot(data = PeterPaul_Tidy,
                  aes(x =month, y = tn_ug, color = lakename)) +
  geom_boxplot() +
  labs(x = "Month", y = "Total N (ug)")

#combine plots into grid
plot_grid(box_temp_clean, box_TP_clean, box_TN_clean, ncol = 1,
          align = 'v', axis = "lr", rel_heights = c(1.2,1,1.3))
```

```
## 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).
```

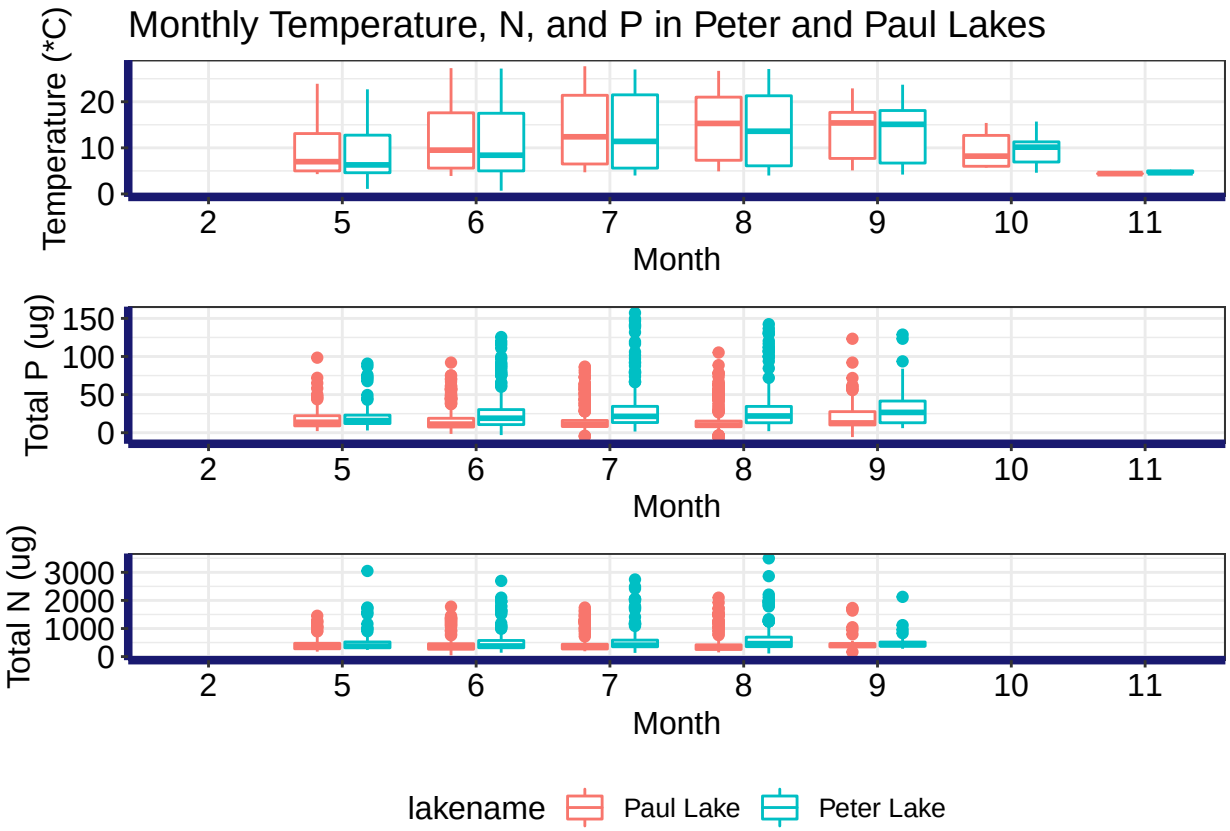

Figure 5: Temperature, total nitrogen, and total phosphorus by month for Peter and Paul lakes

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

> Answer: Temperature has a clear seasonal trend, peaking in the summer (July and August) and reaching a minimum in the late fall (November). Winter data is not available. The temperatures for Peter and Paul lakes are very similar throughout the year. Total nitrogen and total phosphorus are both higher in Peter Lake than Paul Lake. Both TP and TN appear to increase slightly between May and September in Peter Lake, but there is no strong seasonal trend for either in Paul Lake.

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
#create scatterplot of needle dry mass with NLCD color
Needle_Plot <-
  ggplot(subset(Niwot, functionalGroup = "Needles"),
```

```
        aes(x = collectDate, y = dryMass, color = nlcdClass)) +
  geom_point() +
  scale_x_date(date_breaks="3 months") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Date", y = "Dry Mass (g)",
       title = "Dry mass of needles over time, by land use")
print(Needle_Plot)
```

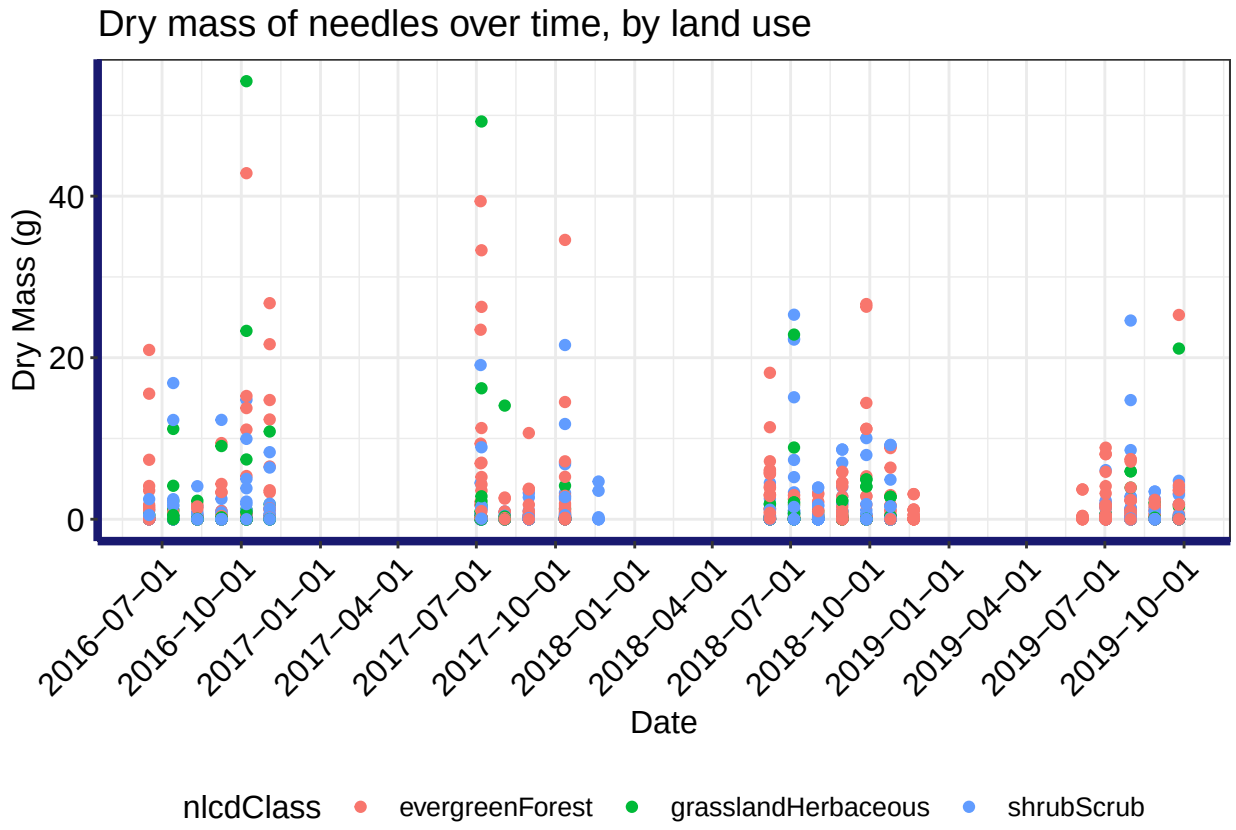

Figure 6: Dry mass of needles by date at Niwot Ridge sites (land cover/land use classes are represented by colors indicated in legend)

```
#7
#create scatterplot of needle dry mass with NCLD facets
Needle_Plot2 <-
  ggplot(subset(Niwot, functionalGroup = "Needles"),
         aes(x = collectDate, y = dryMass)) +
  geom_point() +
  scale_x_date(date_breaks="3 months") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Date", y = "Dry Mass",
       title = "Dry mass of needles over time, by land use") +
  facet_wrap(Niwot$nlcdClass, nrow = 3)
print(Needle_Plot2)
```

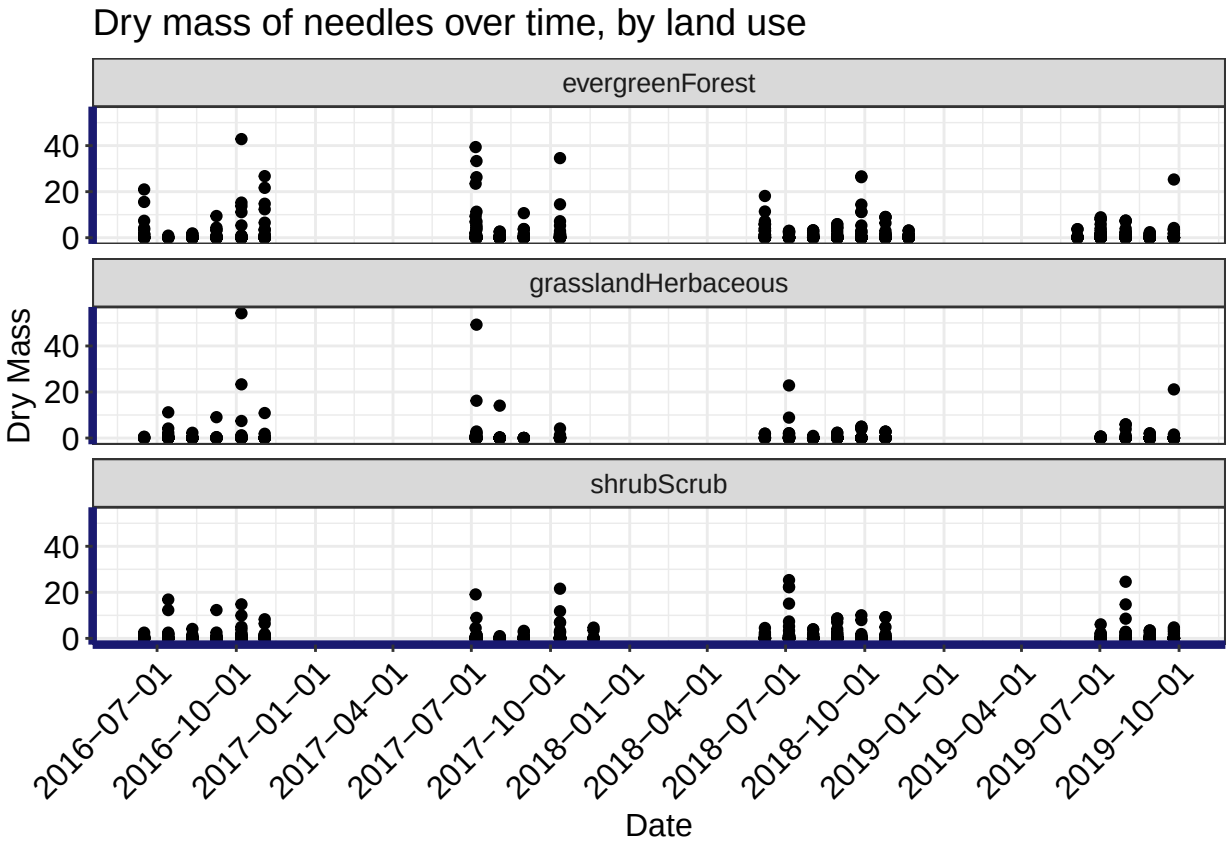

Figure 7: Dry mass of needles by date at Niwot Ridge sites, by land cover/land use class.

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

> Answer: Plot 7 is more effective than plot 6. In plot 7, it is clear that the ranges of dry mass differ for different NLCD classes; overall, the dry mass of needles is lowest in grasslands and highest in evergreen forests. In plot 6, it is clear that the ranges differ for different months - for example, there are generally more instances of high dry mass of needles in july, august, and october - but there are also several low amounts of dry mass for each month, and it is difficult to discern trends between the blue/green/pink dots. Therefore, it is easier to distinguish trends in needle dry mass between land uses when the land uses are represented by facets.

Duke Community Standard affirmation: I have adhered to the Duke Community Standard in completing this assignment. -Anne Harshbarger