Assignment 5: Data Visualization

Katherine Owens

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 Monday, February 14 at 7:00 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 (use the tidy [`NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv`] version) and the processed data file for the Niwot Ridge litter dataset (use the [`NEON_NIWO_Litter_mass_trap_Processed.csv`] version).

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

```
#1
getwd()
```

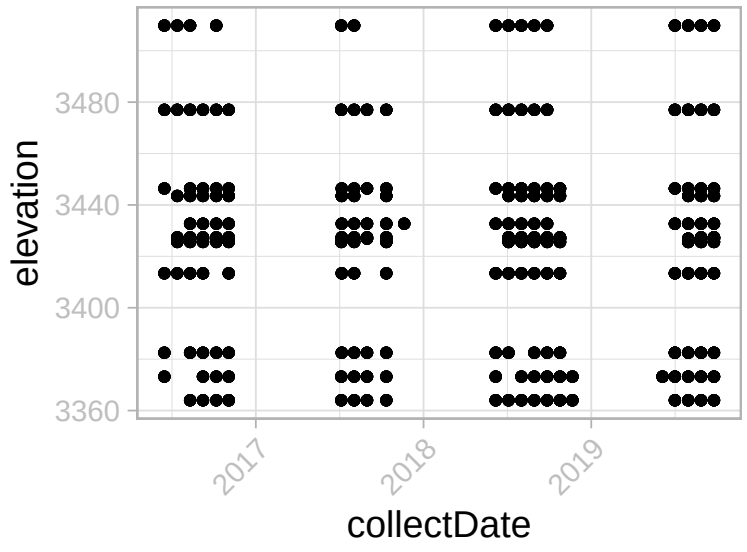
```
## [1] "C:/Users/Katherine/Documents/872-Data Analytics/Environmental_Data_Analytics_2022/Assignments"
```

```
library(tidyverse)
#install.packages("ggridges")
library(ggridges)
library(ggplot2)

PP.ch.nutr <-
  read.csv("../Data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv", stringsFacto
NEON_pr <-
  read.csv("../Data/Processed/NEON_NIWO_Litter_mass_trap_Processed.csv", stringsAsFactors = TRUE)



#2
PP.ch.nutr$sampledate <- as.Date(PP.ch.nutr$sampledate, format = "%Y-%m-%d")

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

Define your theme

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

```
#3 KO version
KOtheme <- theme_light(base_size = 14) +
  theme(axis.text = element_text(color = "gray"),
        legend.position = "top")+ #default legend is on right, moved to top
  theme(axis.text.x = element_text(angle = 45,  hjust = 1))
theme_set(KOtheme)
# options: call the theme in each plot or set the theme at the start.
    #defined own theme as object
    #axis color to gray


KOtest <- ggplot(NEON_pr) +
  geom_point(aes(x = collectDate, y = elevation)) +
  KOtheme
print(KOtest)
```



Create graphs

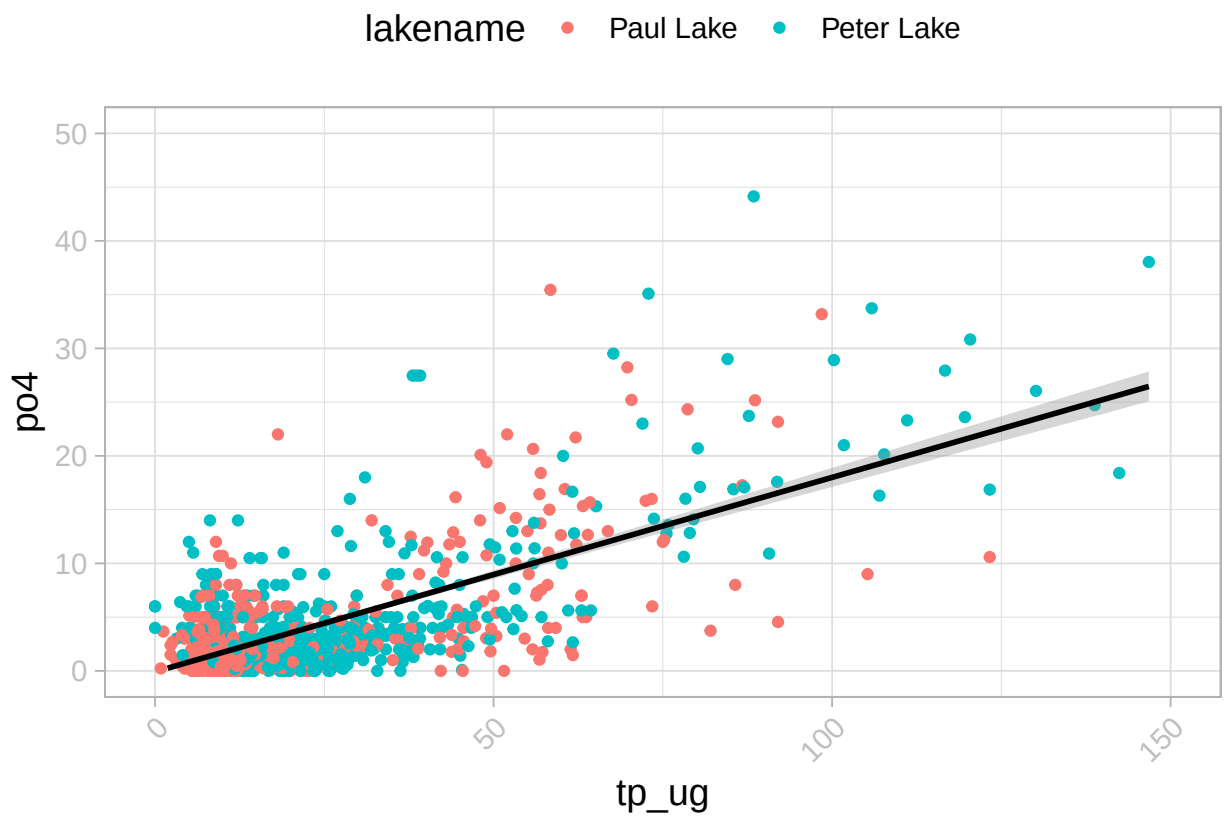
For numbers 4-7, create ggplot graphs and adjust aesthetics to follow best practices for data visualization. Ensure your theme, color palettes, axes, and additional aesthetics are edited accordingly.

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 `ylim()`).

```
#4
tp_po4_P<-
ggplot(PP.ch.nutr, aes(x = tp_ug,
                       y = po4, color = lakename )) +
  geom_point() + #relationship btwn x=tp_ug, y=po4,
                 #categorizing by lakename
  geom_smooth(method = lm, color = "black") +
                 #adding a trendline
```

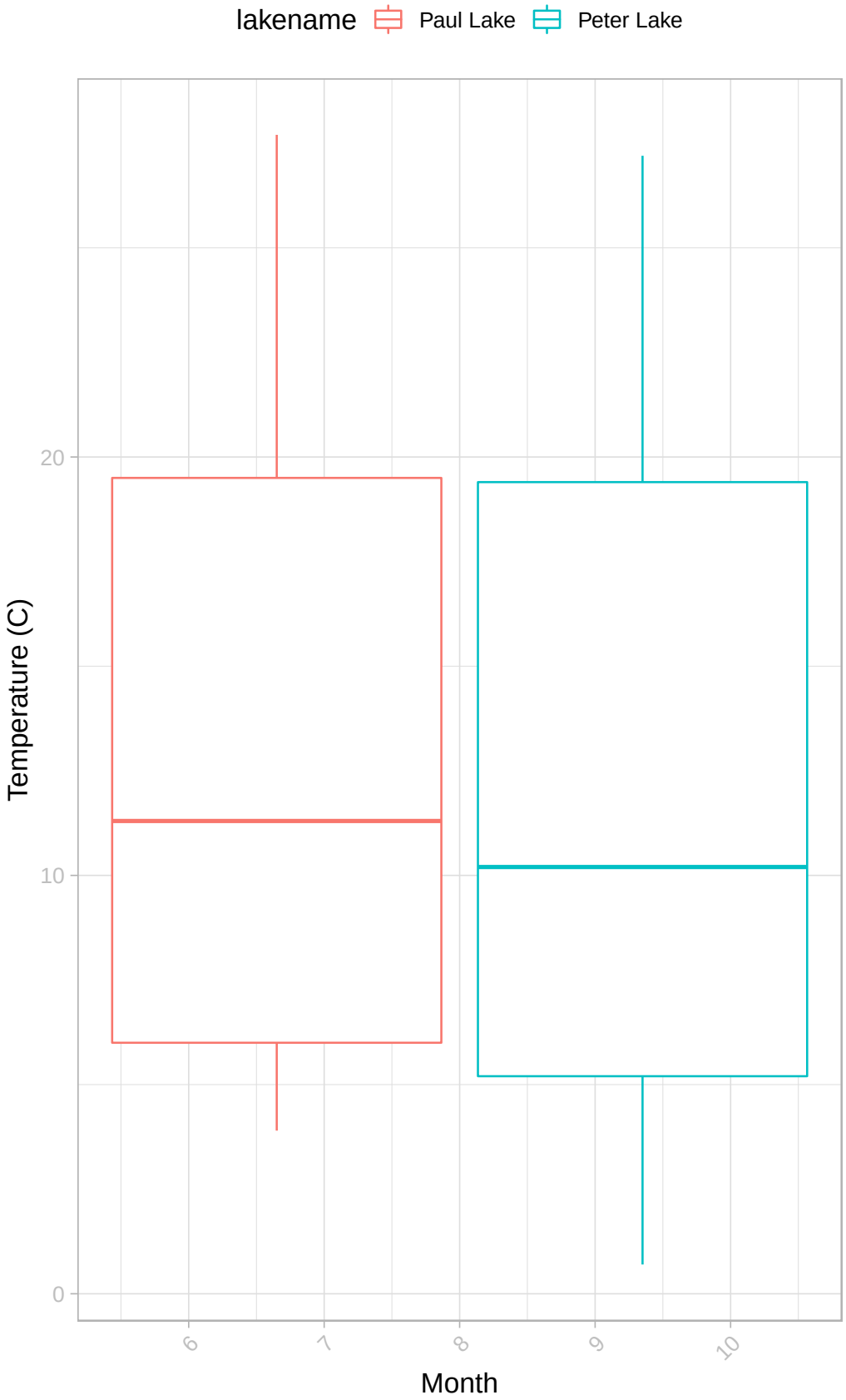
```
  xlim(0, 150) + #excluding three points
  ylim(0, 50)
print(tp_po4_P)
```

```
## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 21948 rows containing non-finite values (stat_smooth).
## Warning: Removed 21948 rows containing missing values (geom_point).
## Warning: Removed 1 rows containing missing values (geom_smooth).
```

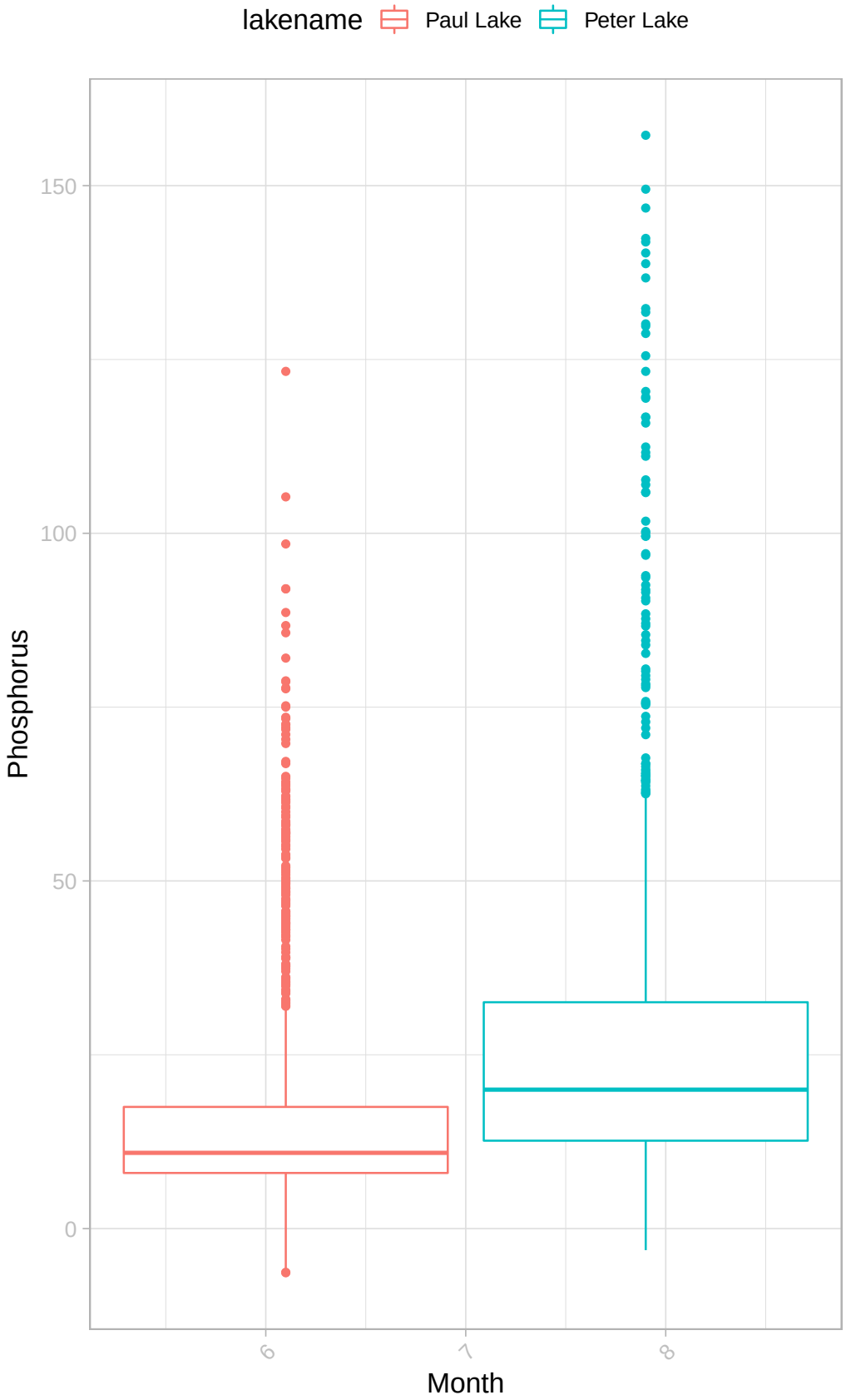


5. [NTL-LTER] Make three separate boxplots of (a) temperature, (b) TP, and (c) TN, with month as the x axis and lake as a color aesthetic. Then, create a cowplot that combines the three graphs. Make sure that only one legend is present and that graph axes are aligned.

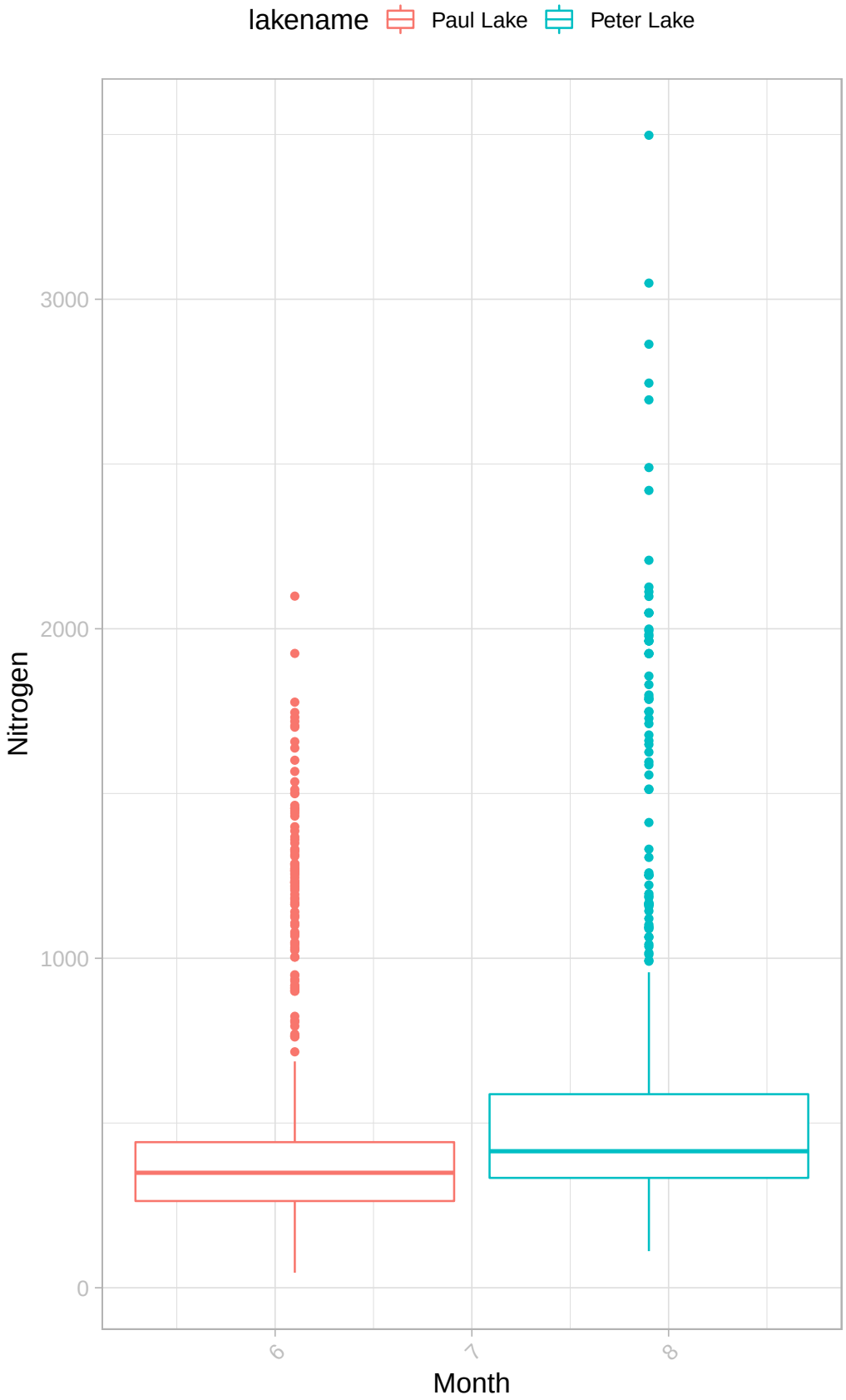
```
#5
#install.packages("cowplot")
library(cowplot)

#Temp plot
Temp_box3 <- ggplot(PP.ch.nutr, aes(x = month,
                    y = temperature_C)) +
  geom_boxplot(aes(color = lakename)) +
  #theme(legend.position = "none") +
    xlab("Month") +
    ylab("Temperature (C)")
print(Temp_box3)
```


lakename
Paul Lake
Peter Lake
Temperature (C)
20
10
0
6
7
8
9
10
Month

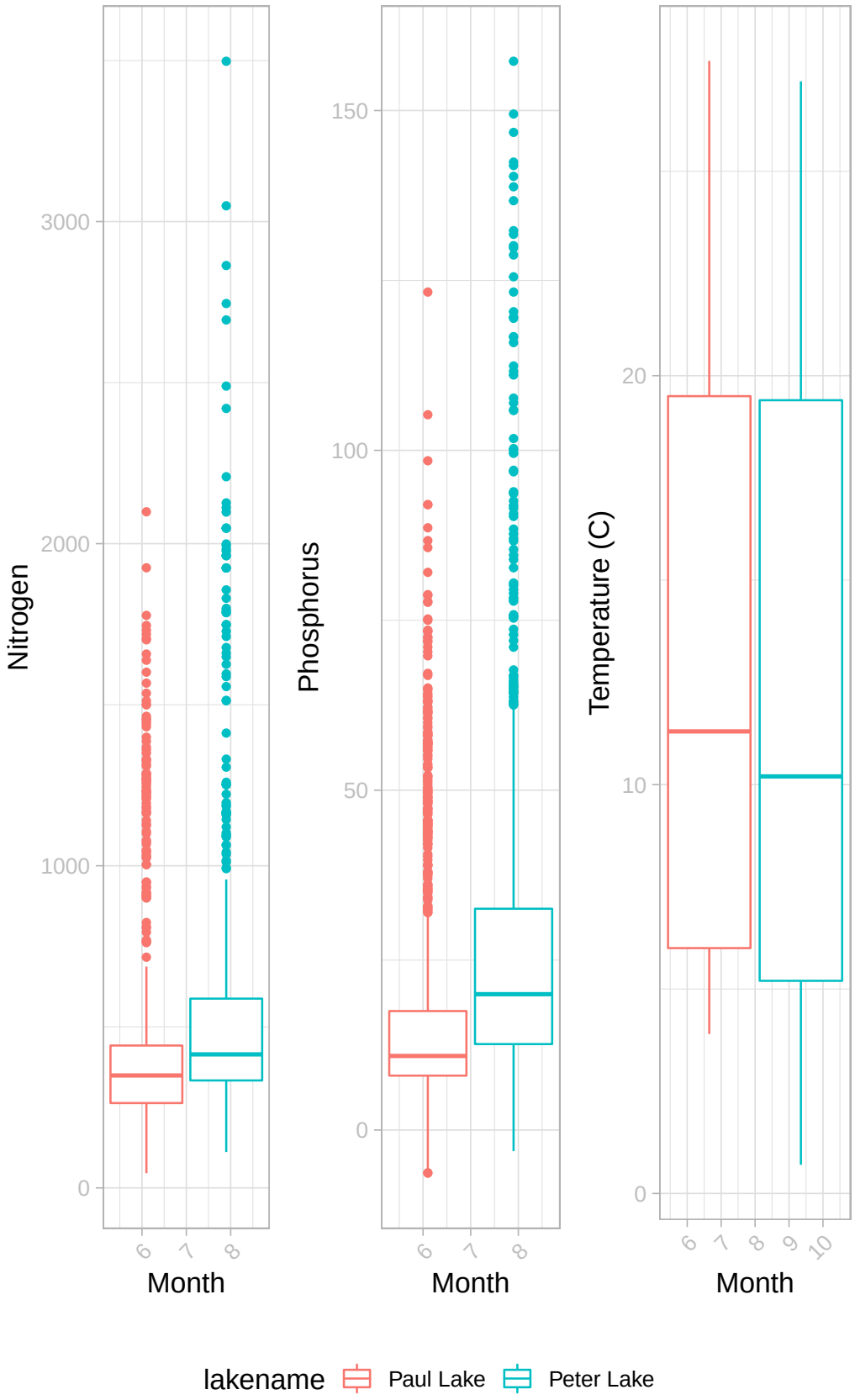
```r
#Phosphorus Plot
TP_box3 <- ggplot(PP.ch.nutr, aes(x = month, y = tp_ug)) +
  geom_boxplot(aes(color = lakename)) +
  #theme(legend.position ="none") +
    xlab("Month") +
    ylab("Phosphorus")
print(TP_box3)
```


lakename
Paul Lake
Peter Lake
150
100
50
0
Phosphorus
6
7
8
Month

```
#Nitrogen Plot
TN_box3 <- ggplot(PP.ch.nutr, aes(x = month, y = tn_ug)) +
  geom_boxplot(aes(color = lakename)) +
  #theme(legend.position="none") +
    xlab("Month") +
    ylab("Nitrogen")
print(TN_box3)
```


lakename
Paul Lake
Peter Lake
3000
2000
1000
0
Nitrogen
6
7
8
Month

```r
#Bringing it all together
legend3 <- get_legend(TN_box3)
PP.ch.nutr$month <- as.factor(PP.ch.nutr$month)
      #month from integer to categorical
Trio_plot3 <- plot_grid(TN_box3 + theme(legend.position = "none"),
                        TP_box3 + theme(legend.position = "none"),
                        Temp_box3 + theme(legend.position = "none"), nrow = 1)


final_plot3 <- plot_grid(Trio_plot3, legend3, nrow = 2,
                         rel_heights = c(3,0.3),
                         align = "v")
print(final_plot3)
```


Nitrogen
3000
2000
1000
0
6
7
8
Month
Phosphorus
150
100
50
0
6
7
8
Month
Temperature (C)
20
10
0
6
7
8
9
10
Month
lakename
Paul Lake
Peter Lake

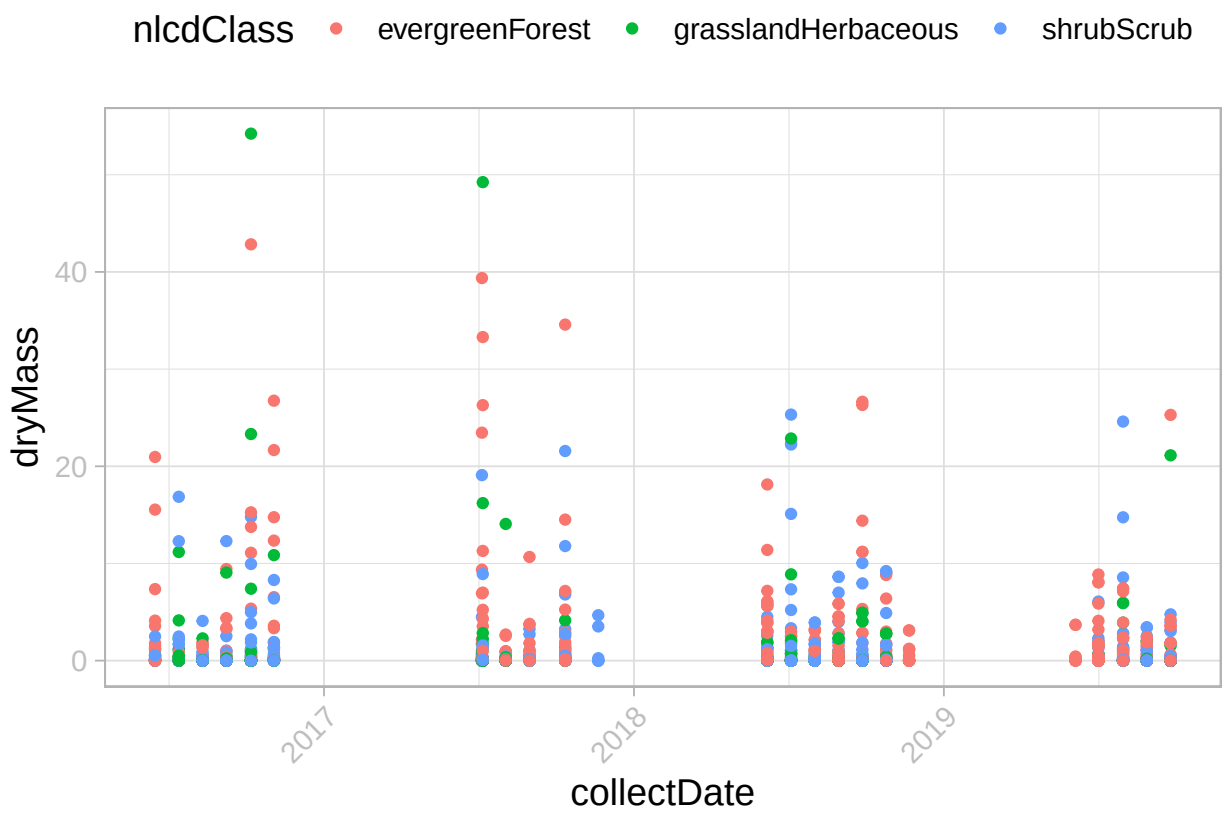
Question: What do you observe about the variables of interest over seasons and between lakes?

> Answer: There seems to be missing data from October-April most of the variables, which may be due to a sampling season trend where collection only happens in the warmer times of the year. Perhaps it has something to do with the amount of sunlight available in the summer versus winter. Also, the Phosphorus and Nitrogen observations are much more spread out with a large range of readings. When compared to temperatures, it is apparent there are less options for possible temperatures with most of the observations concentrating around the two center standard deviations around the mean.
>
> Peter Lake in green has higher levels of phosphorusand nitrogren in general than Paul Lake in orange. Paul lake does have slightly higher average temperature values however, perhaps due to local climate, feeder bodies of water, or higher elevation. Regarding nutrient concentrations, though, maybe Paul Lake is cleaner in general with lower Phosphorus and Nitrogen levels.

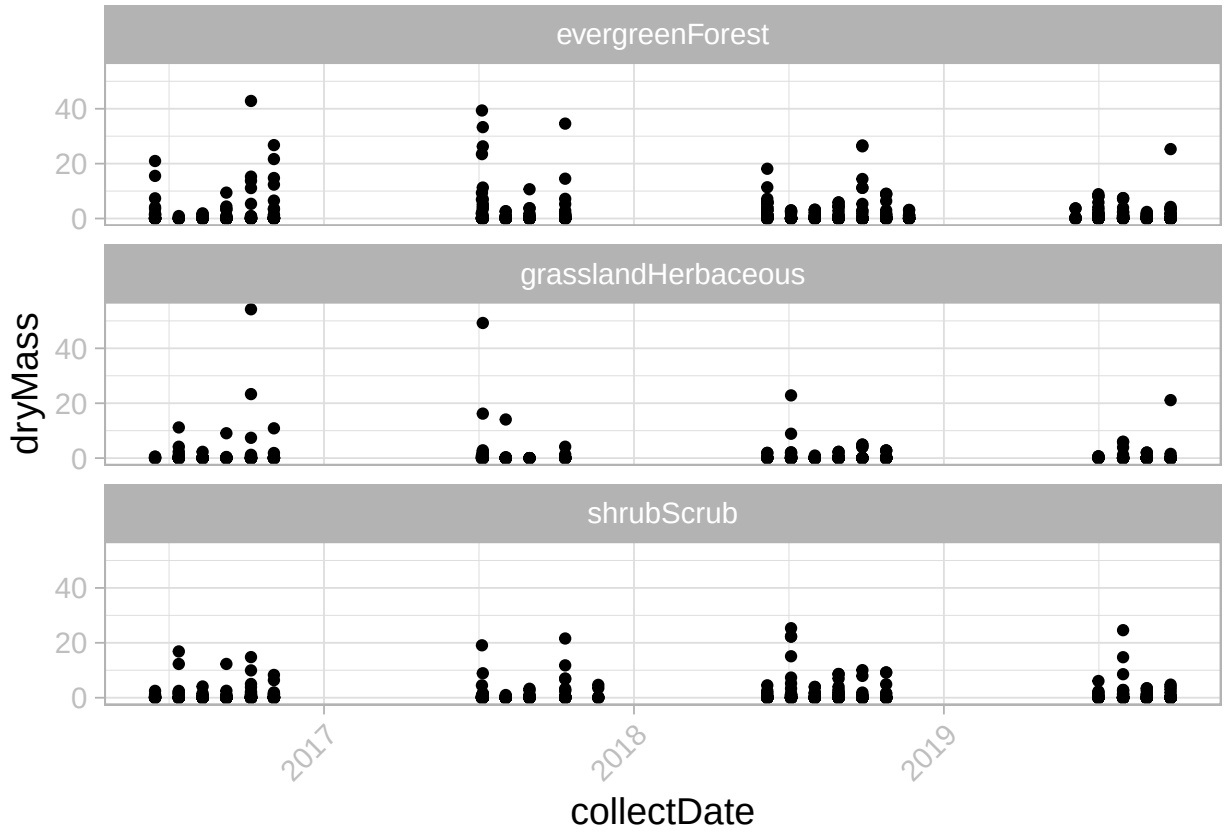
6. [Niwot Ridge] Plot a subset of the litter dataset by displaying only the "Needles" functional group. Plot the dry mass of needle litter by date and separate by NLCD class with a color aesthetic. (no need to adjust the name of each land use)

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

```
#6
drymass_plot1 <-
  ggplot(subset(NEON_pr, functionalGroup = "Needles"), aes(x = collectDate, y = dryMass)) +
  geom_point(aes(color = nlcdClass))
print(drymass_plot1)
```



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



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

> Answer: The second one for number 7 is easier to read because there are no overlaying points, meaning the it is easier to see the trends between the three nlcd classes. Plus the colors from number six don't mean much and are therefore extra noise that is unnecessary.