Assignment 5: Data Visualization

John Rooney

Spring 2023

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A05_DataVisualization.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 your code is tidy; use line breaks to ensure your code fits in the knitted output.
5. Be sure to **answer the questions** in this assignment document.
6. When you have completed the assignment, **Knit** the text and code into a single PDF file.

Set up your session

1. Set up your session. Load the tidyverse, lubridate, here & cowplot packages, and verify your home directory. 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
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(lubridate)
```

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

```r
library(here)
```

```
## here() starts at /Users/jrooney/Library/Mobile Documents/com~apple~CloudDocs/EDA_Sp23/EDA_Sp23
```

```r
library(cowplot)
```

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

```r
getwd()
```

```
## [1] "/Users/jrooney/Library/Mobile Documents/com~apple~CloudDocs/EDA_Sp23/EDA_Sp23"
```

```r
processed_data = "Data/Processed_KEY"

PeterPaul.chem.nutrients <- read.csv(
  here(processed_data,"NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv"),
  stringsAsFactors = T)

Niwot.Ridge.litter <- read.csv(
  here(processed_data, "NEON_NIWO_Litter_mass_trap_Processed.csv"),
  stringsAsFactors = T)

#2
PeterPaul.chem.nutrients$sampledate <- ymd(PeterPaul.chem.nutrients$sampledate)

Niwot.Ridge.litter$collectDate <-
  ymd(Niwot.Ridge.litter$collectDate)
```

Define your theme

3. Build a theme and set it as your default theme. Customize the look of at least two of the following:

- Plot background
- Plot title
- Axis labels

- Axis ticks/gridlines
- Legend

```
#3
my.theme <- theme_light(base_size = 18)+
  theme(axis.text = element_text(color = "grey19"),
        legend.position = "top",
        legend.justification = "left")
```

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.

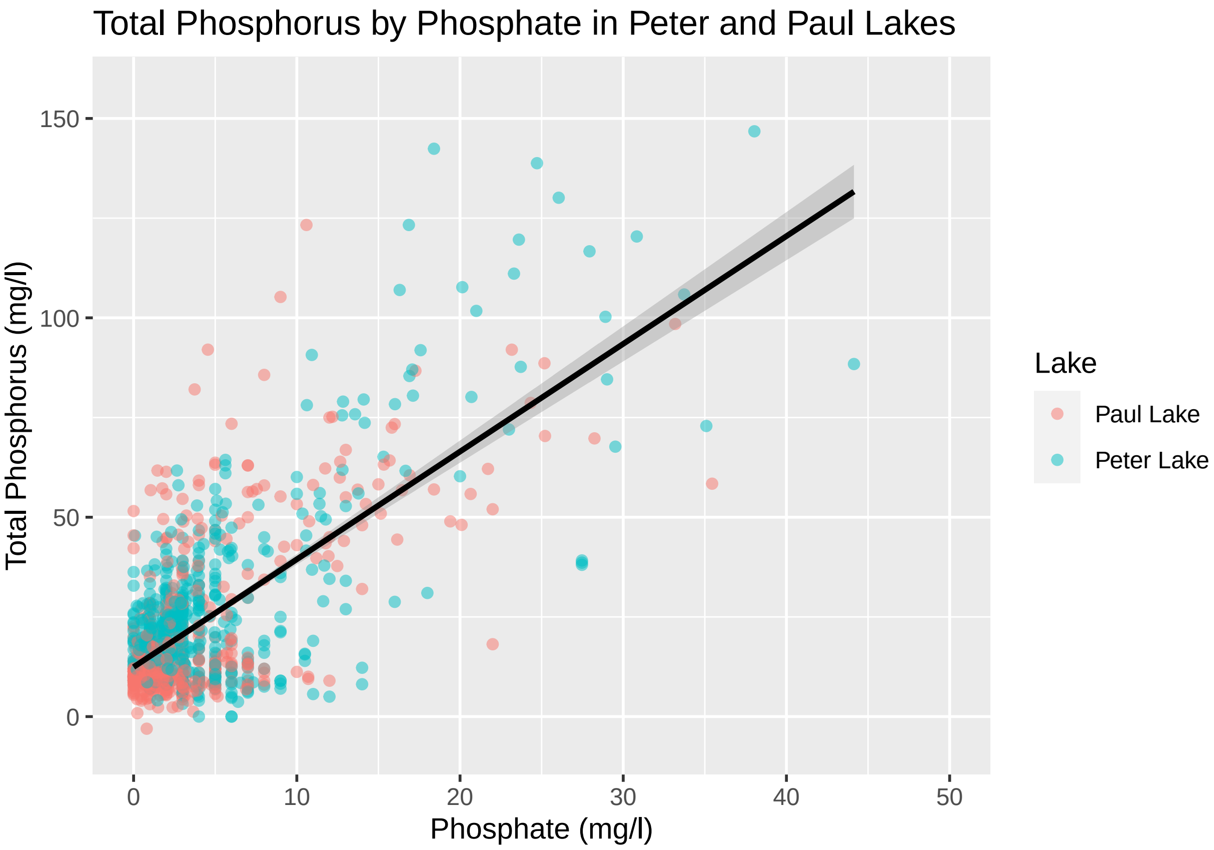
4. [NTL-LTER] Plot total phosphorus (`tp_ug`) by phosphate (`po4`), with separate aesthetics for Peter and Paul lakes. Add a line of best fit and color it black. Adjust your axes to hide extreme values (hint: change the limits using `xlim()` and/or `ylim()`).

```
#4
ggplot(PeterPaul.chem.nutrients,
       aes(x=po4, y=tp_ug, color=lakename))+
         geom_point(,alpha=0.5) +
  geom_smooth(method = lm, color="black")+
  xlim(0,50)+
  labs(x="Phosphate (mg/l)", y="Total Phosphorus (mg/l)", title="Total Phosphorus by Phosphate in Peter
```

```
## ‘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()‘).
```

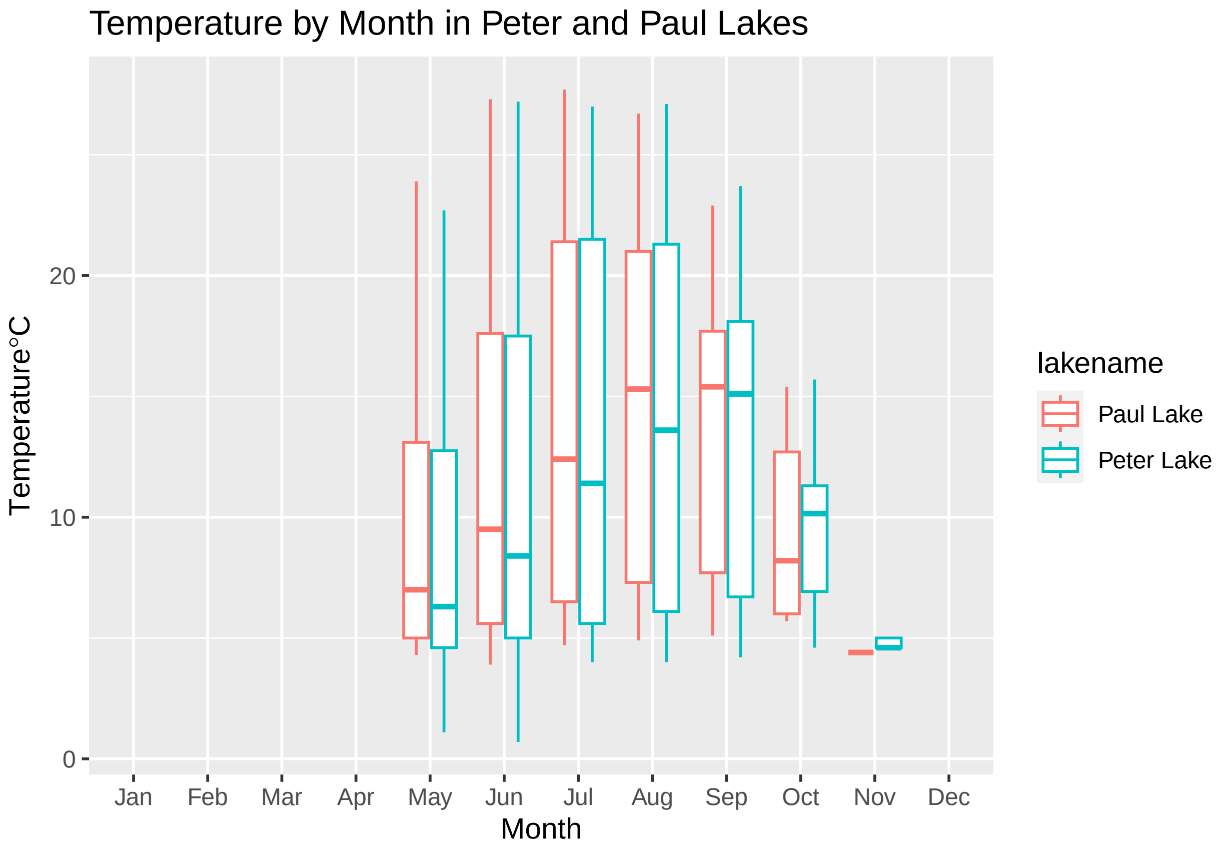


5. [NTL-LTER] Make three separate boxplots of (a) temperature, (b) TP, and (c) TN, with month as the x axis and lake as a color aesthetic. Then, create a cowplot that combines the three graphs. Make sure that only one legend is present and that graph axes are aligned.

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

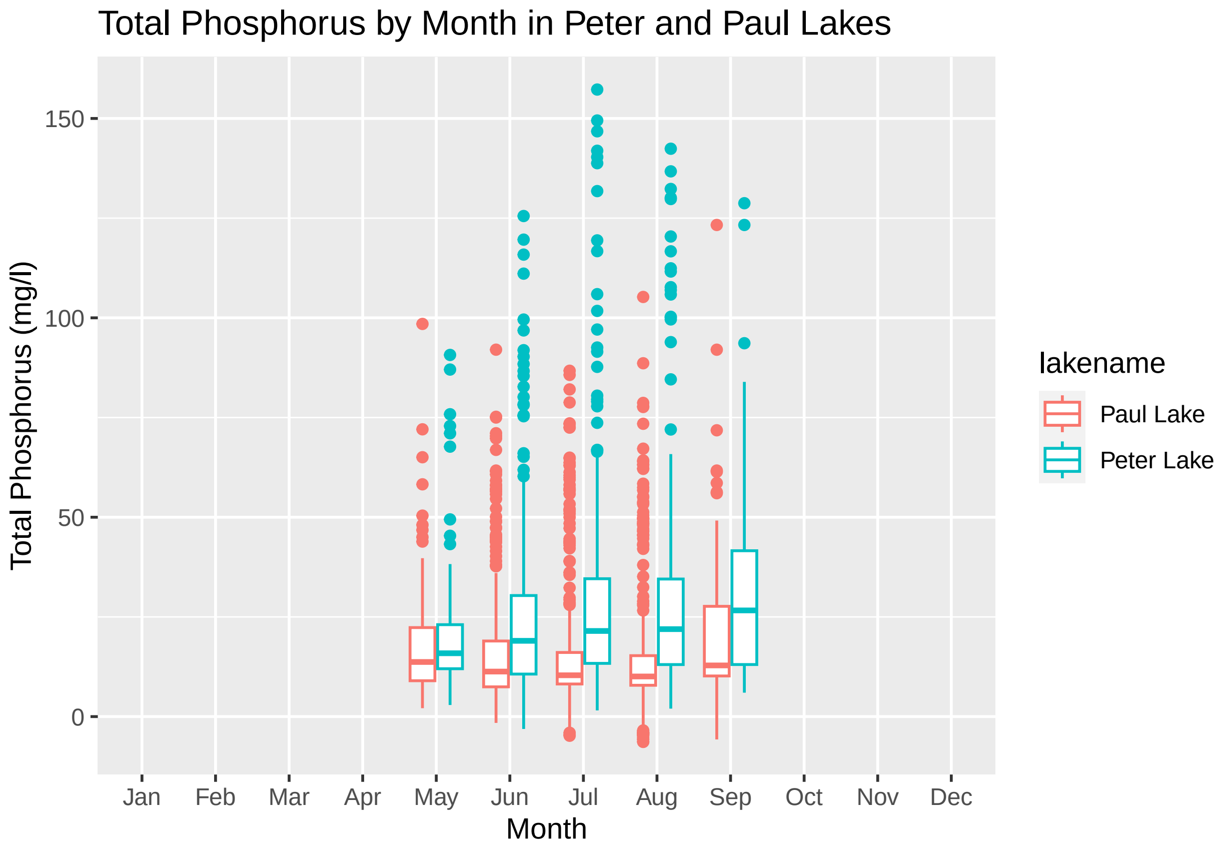
```
#5
boxplot1 <- PeterPaul.chem.nutrients %>%
  ggplot(
    aes(x=factor(
      month,
      levels=1:12,
      labels=month.abb),
      y=temperature_C,
      color=lakename))+
  geom_boxplot()+
  scale_x_discrete(
    name="Month",
    drop=F)+
  labs(y=expression(Temperature*degree*C), title = "Temperature by Month in Peter and Paul Lakes")
print(boxplot1)
```

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



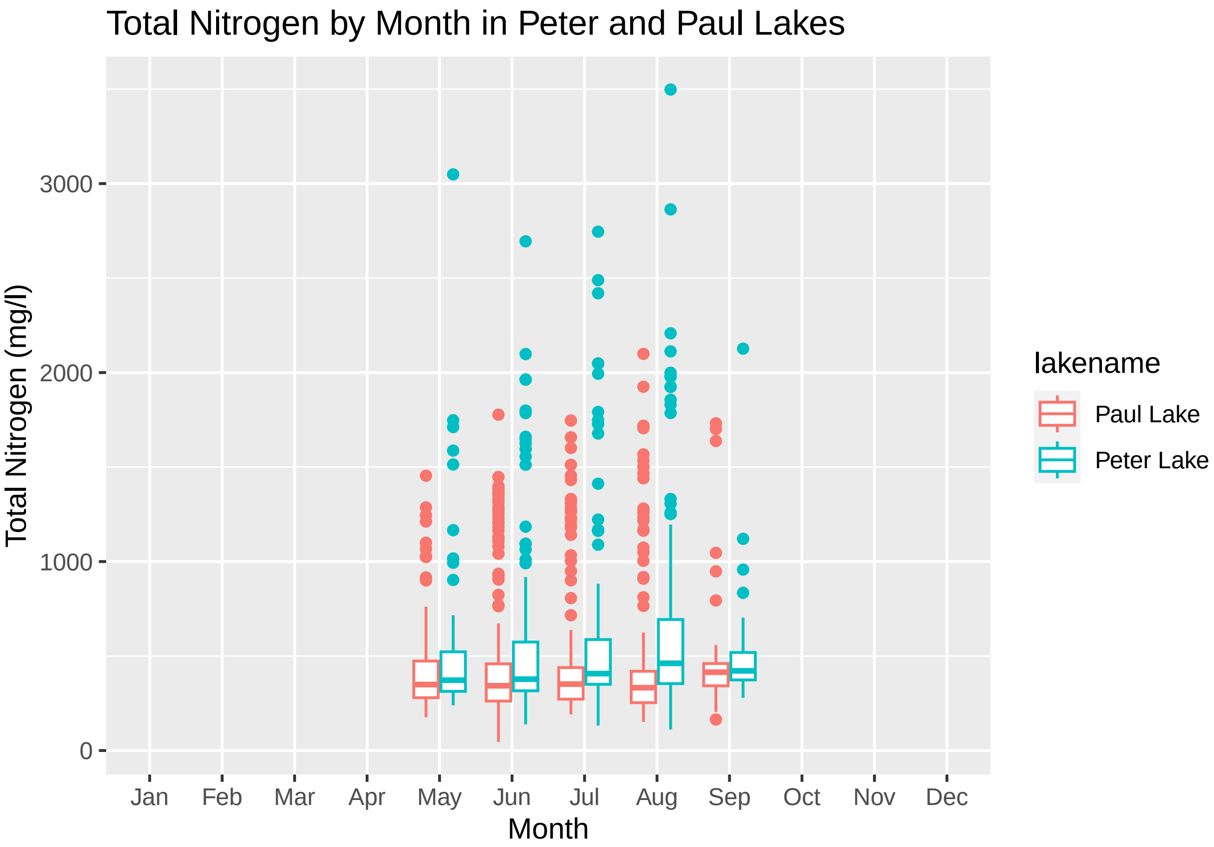
```
boxplot2 <- PeterPaul.chem.nutrients %>%
  ggplot(
    aes(x=factor(
      month,
      levels=1:12,
      labels=month.abb),
      y=tp_ug,
      color=lakename))+
  geom_boxplot()+
  scale_x_discrete(
    name="Month",
    drop=F)+
  labs(y="Total Phosphorus (mg/l)", title = "Total Phosphorus by Month in Peter and Paul Lakes")
print(boxplot2)
```

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



```
boxplot3 <- PeterPaul.chem.nutrients %>%
  ggplot(
    aes(x=factor(
      month,
      levels=1:12,
      labels=month.abb),
      y=tn_ug,
      color=lakename))+
  geom_boxplot()+
  scale_x_discrete(
    name="Month",
    drop=F)+
  labs(y="Total Nitrogen (mg/l)", title = "Total Nitrogen by Month in Peter and Paul Lakes")
print(boxplot3)
```

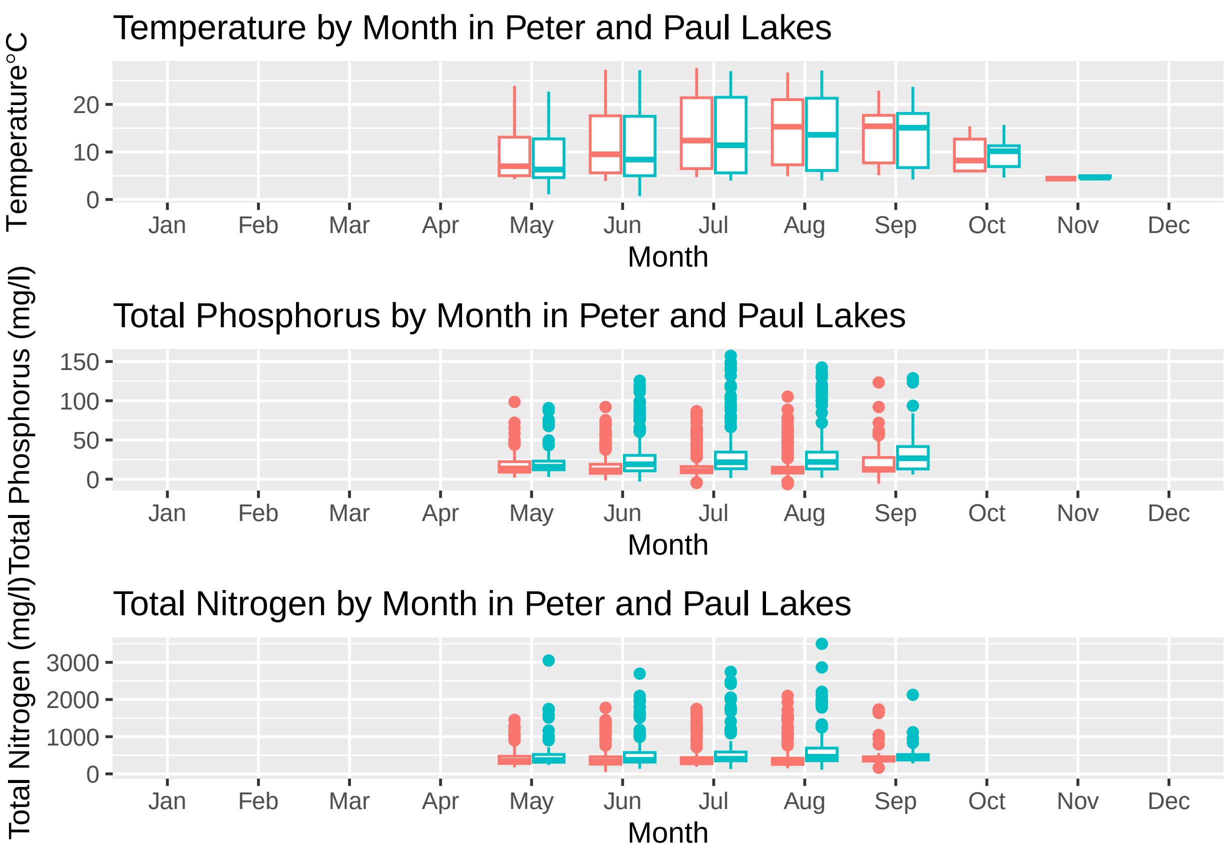
```
## Warning: Removed 21583 rows containing non-finite values (‘stat_boxplot()‘).
```



```
cowplot <- plot_grid(boxplot1 + theme(legend.position = "none"),
                     boxplot2 + theme(legend.position = "none"),
                     boxplot3+ theme(legend.position = "none"),
                     align = "vh",
                     ncol = 1)
```

```
## 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()‘).
```

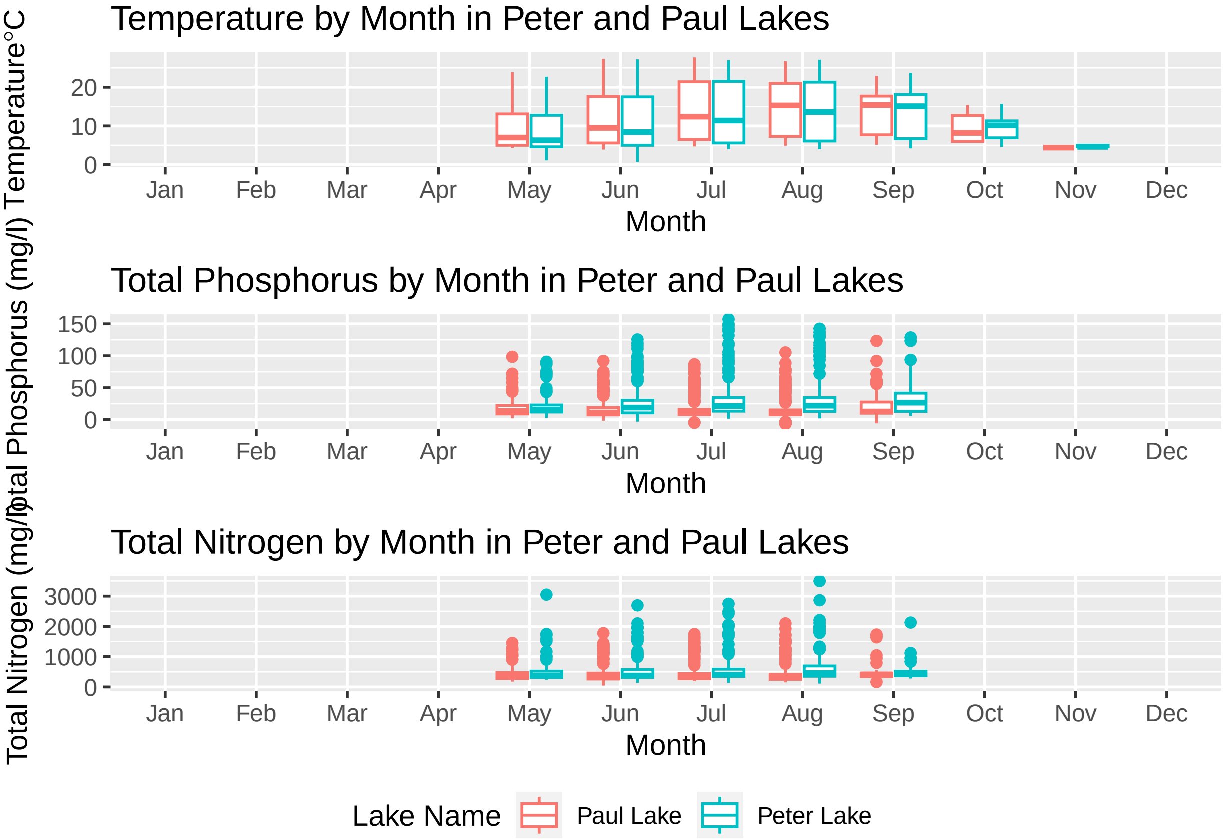
```
cowplot
```



```
legend <- get_legend(boxplot1 +
                       guides(color = guide_legend(nrow = 1)) +
 guides(color = guide_legend(title = "Lake Name")) +                    theme(legend.position = "botto
```

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

```
cowplot.plus.legend <-plot_grid(cowplot, legend, ncol=1, rel_heights = c(1, 0.1))
cowplot.plus.legend
```

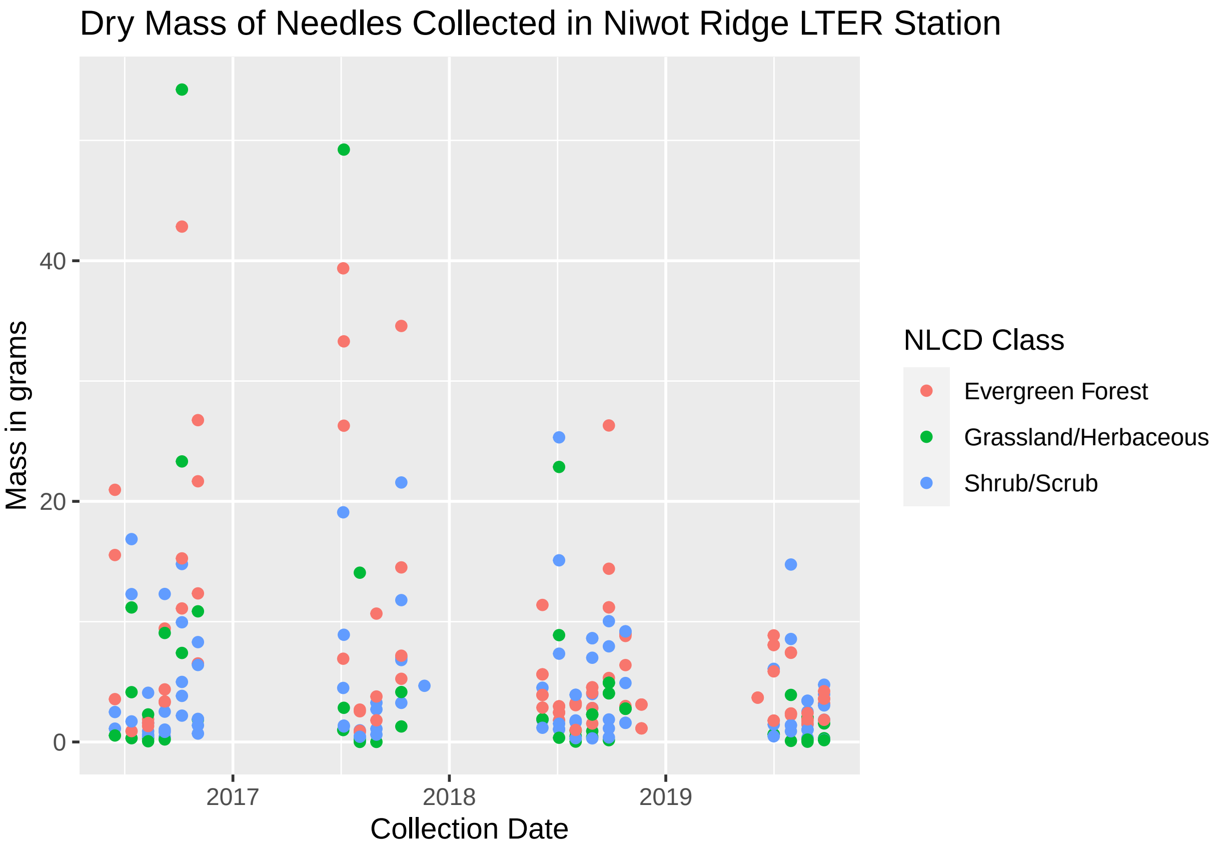


Question: What do you observe about the variables of interest over seasons and between lakes?

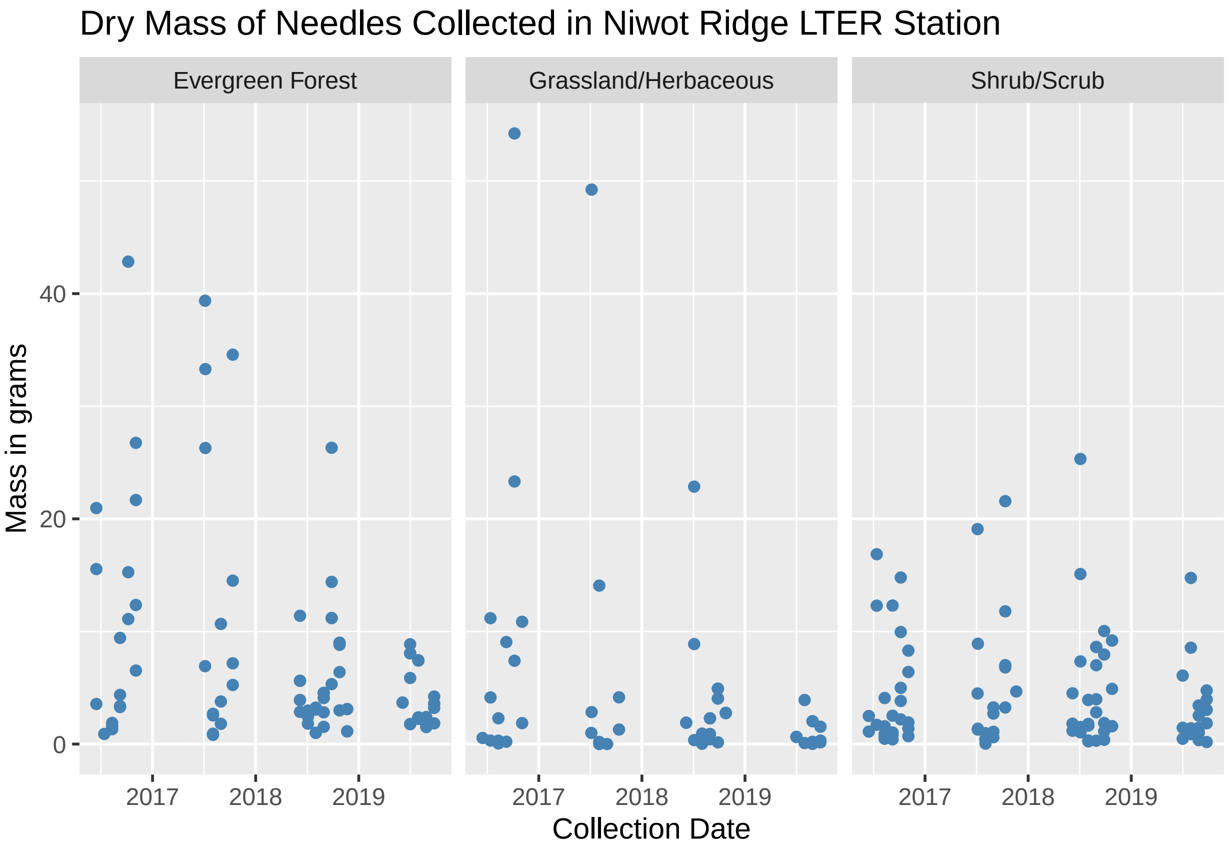
> Answer: The range of the boxplots showing temperature by month look very similar for each lake, with the median for Peter Lake being smaller than Paul Lake in May, June, July, August, and September but then higher in October and November. When looking at total phosphorus, there's more significant outliers for Peter Lake in the summer months, with overall counts in the IQR appearing higher in each month. This pattern holds when looking at nitrogen as well.

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
needles.plot<- Niwot.Ridge.litter %>%
  filter(functionalGroup == "Needles")%>%
  ggplot(aes(x=collectDate, y=dryMass, color=nlcdClass))+
  geom_point()+
  labs(x="Collection Date", y="Mass in grams", title = "Dry Mass of Needles Collected in Niwot Ridge LTE
  scale_color_discrete(labels=c("Evergreen Forest", "Grassland/Herbaceous", "Shrub/Scrub"))

print(needles.plot)
```



```
#7
plot.names<-c('evergreenForest' = "Evergreen Forest",
              'grasslandHerbaceous' = "Grassland/Herbaceous",
              'shrubScrub' = "Shrub/Scrub")

needles.plot.facet <- Niwot.Ridge.litter %>%
  filter(functionalGroup == "Needles")%>%
  ggplot(aes(x=collectDate, y=dryMass)) +
  geom_point(color = "steelblue")+
  labs(x="Collection Date", y="Mass in grams", title = "Dry Mass of Needles Collected in Niwot Ridge LT
  facet_wrap(vars(nlcdClass), labeller=as_labeller(plot.names), ncol = 3)

print(needles.plot.facet)
```



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

Answer: I think the plot produced in Question 7 is more effective as it's easier to see the differences in mass across locations. At least for me, it's easier to see which location had outliers in the plot produced in Question 6, but harder to see patterns by location.