Assignment 5: Data Visualization

Nancy Bao

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
#Verifying working directory
getwd()
```

```
## [1] "/Users/Nancy/Desktop/Semester 4/ENV 872L/Environmental_Data_Analytics_2021"
```

```
#Load necessary packages
#install.packages("wesanderson")
#I installed wesanderson for more colors
library(tidyverse)
library(ggplot2)
library(ggridges)
library(cowplot)
library(viridis)
library(RColorBrewer)
library(colormap)
library(wesanderson)
# I referred to https://cran.r-project.org/web/packages/wesanderson/wesanderson.pdf
#The above URL was used to explore different palettes in the wesanderson package
#Import the data files for Peter and Paul Lakes

NTL_PeterPaul_nutrients<-
```

```
  read.csv("./Data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv"
        , stringsAsFactors = TRUE)
NTL_PeterPaul_nutrients_gathered<-
  read.csv("./Data/Processed/NTL-LTER_Lake_Nutrients_PeterPaulGathered_Processed.csv"
           , stringsAsFactors = TRUE)
NW_Ridge_litter<-read.csv("./Data/Processed/NEON_NIWO_Litter_mass_trap_Processed.csv",
                          stringsAsFactors = TRUE)
#2
##NTL_PeterPaul_nutrients: Change from factor to date
class(NTL_PeterPaul_nutrients$sampledate)
```

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

```
NTL_PeterPaul_nutrients$sampledate<-as.Date(NTL_PeterPaul_nutrients$sampledate,
                                            format = "%Y-%m-%d")
#verifying class is now date
class(NTL_PeterPaul_nutrients$sampledate)
```

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

```
##NTL_PeterPaul_nutrients_gathered: Change from factor to date
class(NTL_PeterPaul_nutrients_gathered$sampledate)
```

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

```
NTL_PeterPaul_nutrients_gathered$sampledate<-
  as.Date(NTL_PeterPaul_nutrients_gathered$sampledate,
          format = "%Y-%m-%d")
#verifying class is now date
class(NTL_PeterPaul_nutrients_gathered$sampledate)
```

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

```
##NW_Ridge_litter: Change from factor to date
class(NW_Ridge_litter$collectDate)
```

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

```
NW_Ridge_litter$collectDate<-as.Date(NW_Ridge_litter$collectDate,
                                     format = "%Y-%m-%d")
#verifying class is now date
class(NW_Ridge_litter$collectDate)
```

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

Define your theme

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

```
# Setting my default theme as hw_theme
hw_theme <-theme_bw(base_size=14)+
      theme(legend.position = "top",
        legend.justification = "center",
        legend.text = element_text(size = 12),
        legend.title = element_text(size = 12,face= "bold"),
        plot.title = element_text(hjust = 0.5,size=14))
# I made my font size 14 with base_size(), I picked the theme_bw() as my default
# I also chose to have my legend at the top and center justified with 12pt font
# I used face=bold to get bold font for the legend title
# I used plot.title = element_text(hjust = 0.5,size=14)) to center my plot titles
theme_set(hw_theme)
```

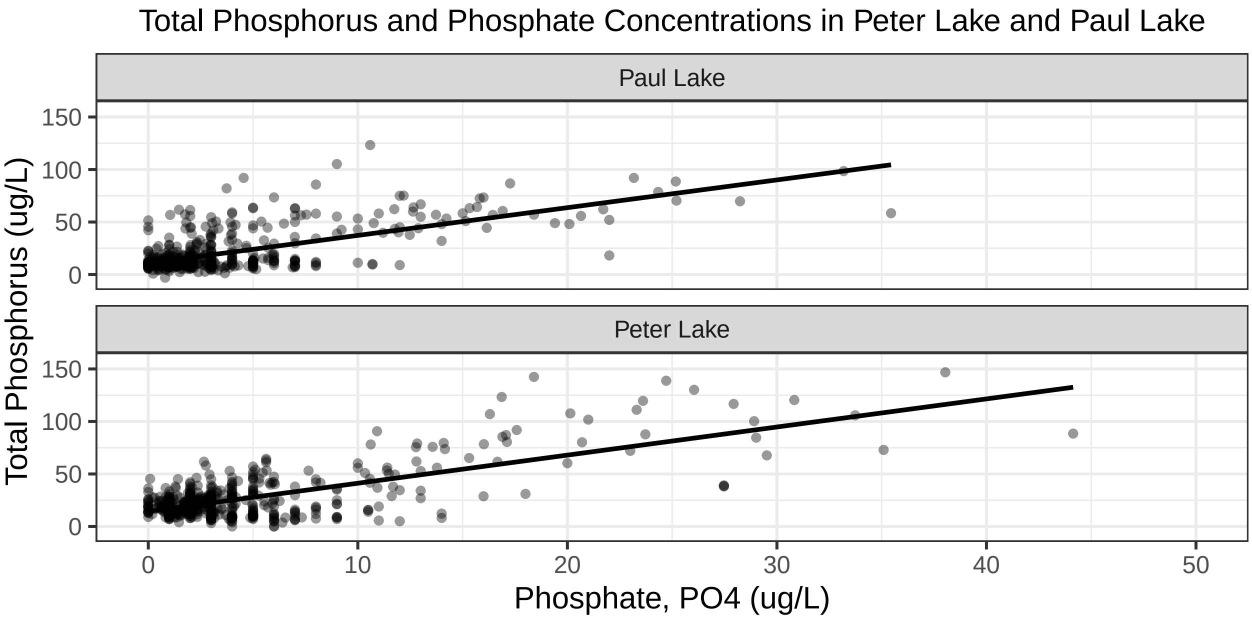
Create graphs

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.

```
#4 Plot tp_ug by po4 with separate aesthetics for Peter and Paul Lakes
#both tp_ug and po4 are continuous variables
#I set tp_ug as the y-axis and po4 as the x-axis
#
TotP_vsPO4_plot<-
  ggplot(NTL_PeterPaul_nutrients,aes(x= po4, y = tp_ug)) +
         geom_point(alpha=0.4, size=1.5) +
         geom_smooth(method=lm,color="black",se=FALSE) +
         facet_wrap(vars(lakename),nrow=2) +
         xlim(0,50) +
         ylab("Total Phosphorus (ug/L)") + #relabeled y axis title
         xlab("Phosphate, PO4 (ug/L)") + #relabeled x axis title
         labs(title="Total Phosphorus and Phosphate Concentrations in Peter Lake and Paul Lake")
print(TotP_vsPO4_plot)
```

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

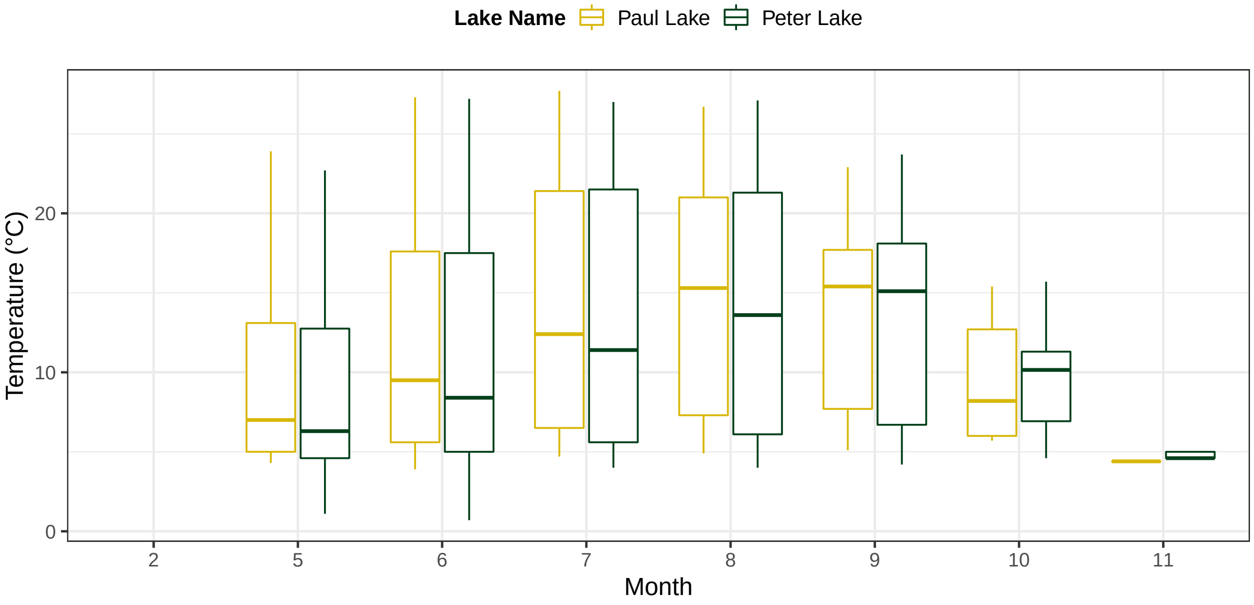


```
#Originally, I had all points on one plot, and differentiated color by factoring by lakename.
#However, that obscured the trends the different lakes.
#I used facet wrap to separate the aesthetics for the two lakes.
#I decided not to change the lakes to different colors
#because they are already labeled by lake, and I wanted to make this inclusive
#I did change alpha to 0.4, so you could still see the trend line
```

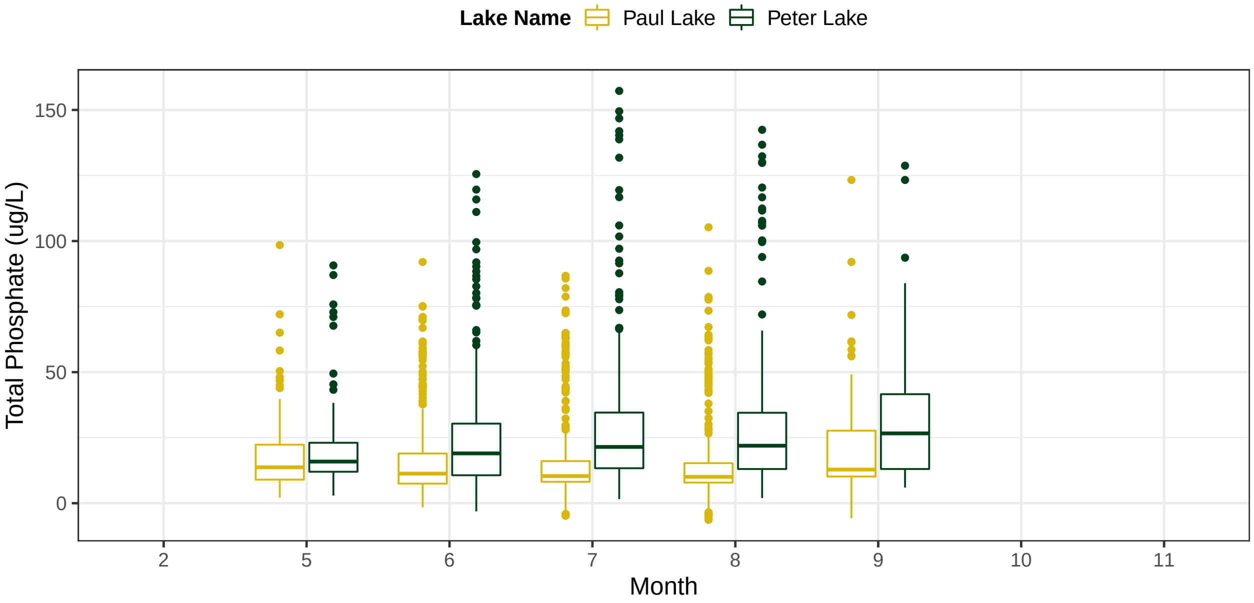
```
#so that those with colorblindness and/or weakness could still see the trends
#I included se=FALSE to remove the confidence interval around the regression line.
#There was one extraneous point in the 300 ug/L range for phosphate conc.
#I used xlim() to created a better spread of data, and exclude the extreme value.
#I set the x limit to 50ug/L 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.

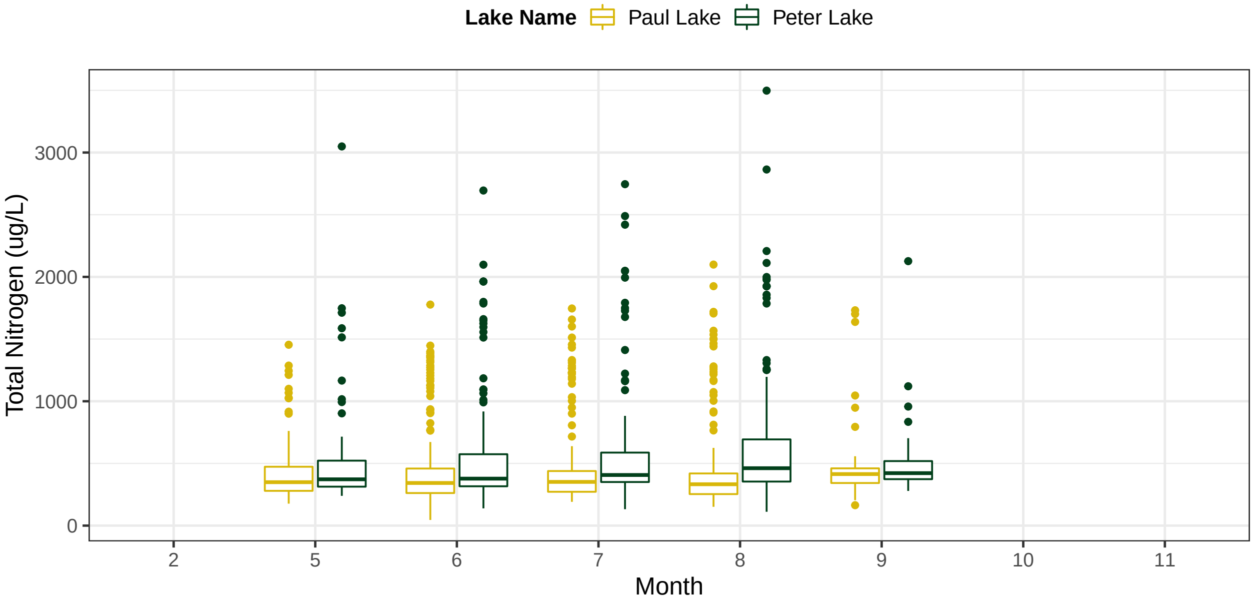
```
#boxplots: x axis is month and lake is color aesthetic
#I used wes_palette("Cavalcanti1") for the boxplot colors from the wesanderson package
#determined nutrient concentrations were ug/L from 05_Part2_DataVisualization
#I changed month from integer to factor with as.factor() in the aes() function
#In order to make a boxplot, month needed to be a categorical variable.
#5a. temperature
NTL_temp_boxplot<-ggplot(NTL_PeterPaul_nutrients, aes(x= as.factor(month),
                  y= temperature_C,color=lakename))+
                  geom_boxplot()+
                  scale_color_manual(values = wes_palette("Cavalcanti1"))+
                  xlab("Month")+ylab("Temperature (°C)")+
                  labs(color="Lake Name")
print(NTL_temp_boxplot)
```



```
#5b. TP
NTL_TP_boxplot<-ggplot(NTL_PeterPaul_nutrients,
                aes(x=as.factor(month), y= tp_ug, color= lakename))+
                geom_boxplot()+
                scale_color_manual(values = wes_palette("Cavalcanti1"))+
                xlab("Month")+
                ylab("Total Phosphate (ug/L)")+
                labs(color="Lake Name")
print(NTL_TP_boxplot)
```

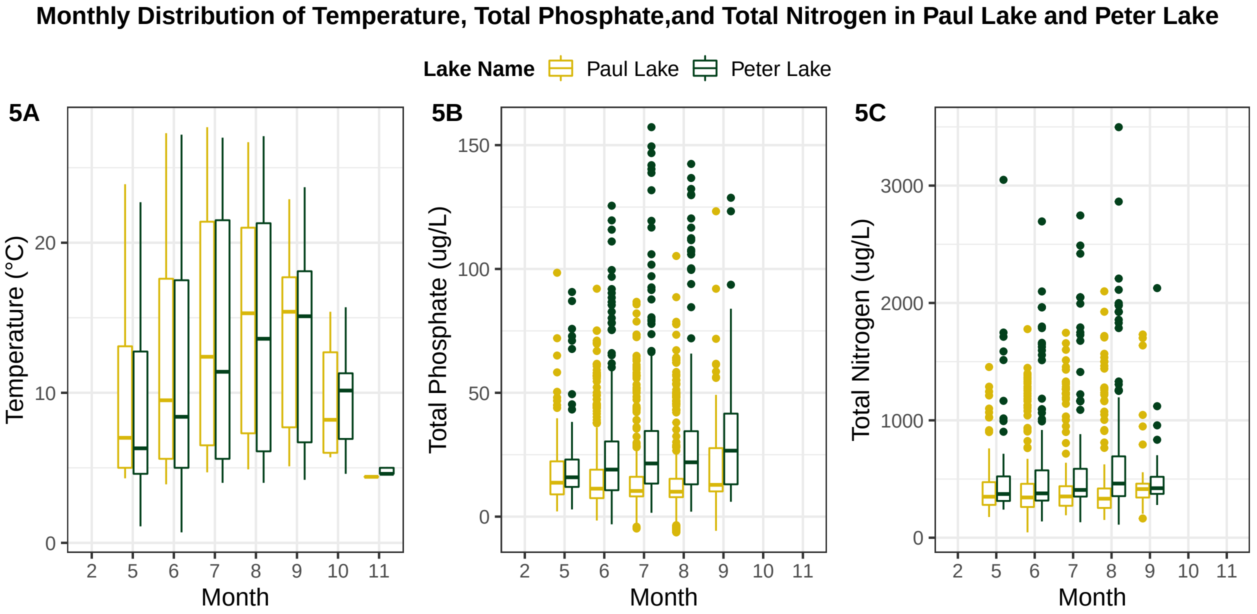


```
#5c. TN
NTL_TN_boxplot<-ggplot(NTL_PeterPaul_nutrients,
                        aes(x=as.factor(month), y=tn_ug, color=lakename))+
                        geom_boxplot()+
                        scale_color_manual(values = wes_palette("Cavalcanti1"))+
                        xlab("Month")+
                        ylab("Total Nitrogen (ug/L)")+
                        labs(color="Lake Name")
print(NTL_TN_boxplot)
```



```
#Cowplot combining graphs from 5a-5c
#https://cran.r-project.org/web/packages/cowplot/cowplot.pdf
#I used URL above to read up on get_legend
#set overall legend with get_legend(NTL_TP_boxplot)
#since all the boxplots have same color code, I just picked any plot
#I used NTL_TP_boxplot to set overall legend
```

```
NW_legend<-get_legend(NTL_TP_boxplot)
combined_5a_c_plot<-plot_grid(NTL_temp_boxplot+theme(legend.position="none"),
                    NTL_TP_boxplot+theme(legend.position="none"),
                    NTL_TN_boxplot+theme(legend.position="none"),
                    nrow=1, align="h",rel_heights = c(2, 1),
                    labels = c("5A", "5B", "5C"))
#I used theme(legend.position="none") to hide the individual legends
#I used labels=c() to rename the graphs so you know which one is a, b, and c
#added a title to plot using ggdraw() and draw_label
#I referred to https://cran.r-project.org/web/packages/cowplot/cowplot.pdf
plot_title <- ggdraw() +
draw_label("Monthly Distribution of Temperature, Total Phosphate,and Total Nitrogen in Paul Lake and Pe
#I used another cowplot to combine the combined boxplots with the NW_legend and title I created above
final_5a_c_plot<-plot_grid(plot_title,NW_legend,combined_5a_c_plot,ncol=1,rel_heights = c(0.1,0.1,1))
#rel_heights=c() was used to adjust distance of title from legend from plots
print(final_5a_c_plot)
```



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

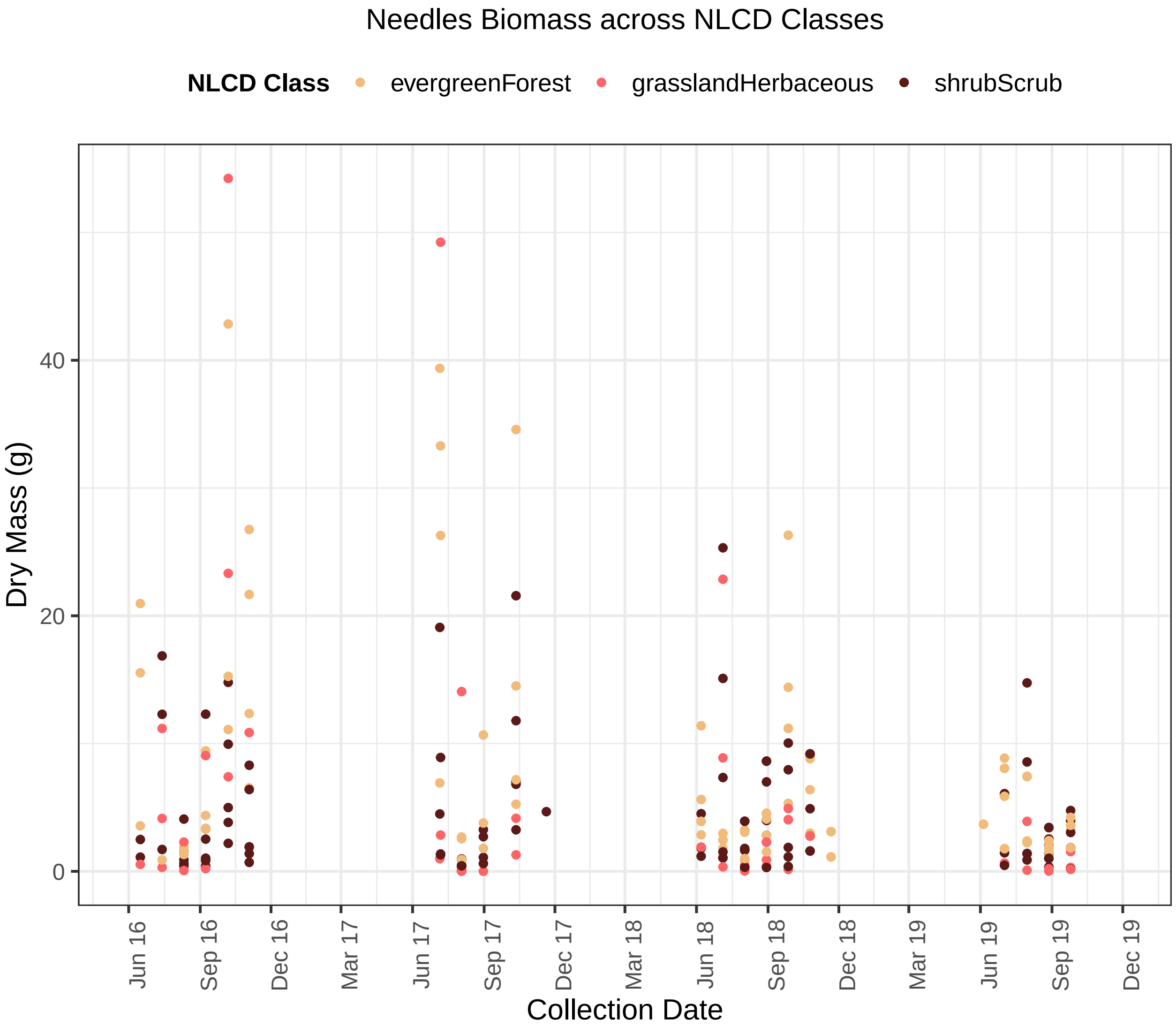
> Answer: From the graph developed in question 4, I see a positive correlation between total phosphorus concentration and phosphate concentration for both Paul Lake and Peter Lake. The values are mainly clusted between 0ug/L and 15 ug/L for both Paul Lake and Peter Lake. However, Peter Lake has points that are more spread out towards the higher phosphate concentrations between 20 and 45ug/L. Temperature increases from May to August and Early September, which makes sense as we move from spring to summer and there is more sunlight to warm the water (Fig 5A). The temperature drops in October and November, which follows the trend for the fall in the Northern hemisphere. The Total Phosphate concentrations and total nitrogen concentrations have many outliers that skew the data to the right. The total phosphate and total nitrogen concentrations generally are higher in Peter Lake than Paul Lake (Fig 5B). Total phosphate concentrations seem to increase from spring to summer months for Peter Lake, while they stay consistent and slightly dip for Paul Lake. The total nitrogen concentrations are consistent across the spring and summer months for both lakes (Fig 5C).

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)

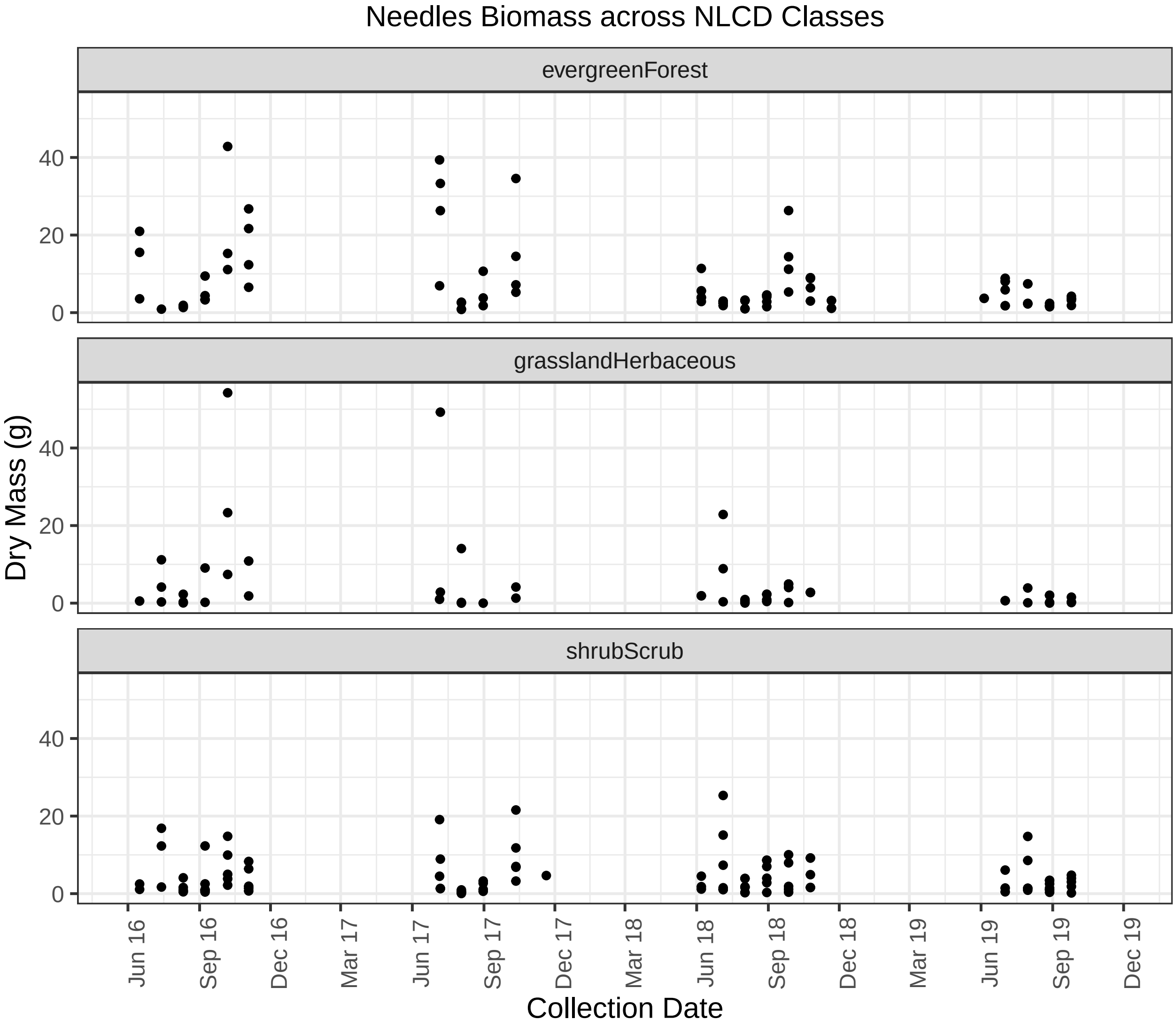
7. [Niwot Ridge] Now, plot the same plot but with NLCD classes separated into three facets rather than separated by color.

```
#6
Needles.NW.plot<-ggplot(subset(NW_Ridge_litter, functionalGroup== "Needles"),
                              aes(x= collectDate, y= dryMass, color= nlcdClass))+
                              geom_point()+
                              scale_x_date(limits = as.Date(c("2016-05-31", "2019-11-30")),
                              date_breaks = "3 months", date_labels = "%b %y")+
                              xlab("Collection Date")+
                              ylab("Dry Mass (g)")+
                              labs(color="NLCD Class")+
                              labs(title="Needles Biomass across NLCD Classes")+
                              theme(axis.text.x = element_text(angle = 90))+
                              scale_color_manual(values=wes_palette("GrandBudapest1"))
print(Needles.NW.plot)
```



```
#I used date_labels =%b %y to show date as month and year rather than just year to assess season
# For question 6, I decided to use the wesanderson package and use the GrandBudapest palette
```

```
#I thought the GrandBudapest1 palette had more contrasting value between the colors than the default.
#I used https://cran.r-project.org/web/packages/wesanderson/wesanderson.pdf to find palette
#I chose a scatterplot because dry mass and collection date are continuous variables
#7
Needles.NW.plot.facet<- ggplot(subset(NW_Ridge_litter, functionalGroup== "Needles"),
                               aes(x= collectDate, y= dryMass))+
                               geom_point()+
                               facet_wrap(vars(nlcdClass),nrow=3)+
                               scale_x_date(limits = as.Date(c("2016-05-31", "2019-11-30")),
                               date_breaks = "3 months", date_labels = "%b %y")+
                               theme(axis.text.x = element_text(angle = 90))+
                               xlab("Collection Date")+
                               ylab("Dry Mass (g)")+
                               labs(title="Needles Biomass across NLCD Classes")
#For plots made in question 6 and 7
#I changed the date range to fit the collection date range
#I added date+labels by month and year to make it easier to differentiate seasons
#I used date_breaks of 3 months
#3 months provided intervals to assess seasonality differences
#I rotated the x axis text because the text of the labels were overlapping using the following:
#theme(axis.text.x = element_text(angle = 90))
print(Needles.NW.plot.facet)
```



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

Answer: I think plot 7 is more effective because I can clearly see the seasonal dry needles mass differences for each NLCD class. With plot 6, all the land covers are put on the same graph and some of the points over lap. Furthermore, the facet has one color for all the graphs, but is clearly labeled. The colors in plot 6 may confuse some viewers and is not as inclusive (we need to consider being more inclusive for those viewing these plots who have color blindness and color weakness) compared to the plot in question 7. The trend is clearer with the plot from question 7. I see that dry biomass mass slightly decreases across the summer collection dates from 2016 to 2019, which was not as clear with the plot from question 6. I can also see the spread of dry needle biomass mass has a greater range for evergreen forest versus shrubScrub. This makes sense as evergreen forests are more likely to be inhabitated by coniferous species that have needles than grasslands.