Assignment 5: Data Visualization

Yikai Jing

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] "/Users/me/Environmental_Data_Analytics_2022/Assignments"
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
```

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

```
library(cowplot)
NCP <- read.csv("~/Environmental_Data_Analytics_2022/Data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_P
NIL <- read.csv("~/Environmental_Data_Analytics_2022/Data/Processed/NEON_NIWO_Litter_mass_trap_Processe
#2
class(NCP$sampledate)
```

```
## [1] "character"
```

```
NCP$sampledate <- as.Date(NCP$sampledate)
class(NCP$sampledate)
```

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

```
class(NIL$collectDate)
```

```
## [1] "character"
```

```
NIL$collectDate <- as.Date(NIL$collectDate)
class(NIL$collectDate)
```

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

Define your theme

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

```
#3
theme_set(theme_classic())
```

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()`).

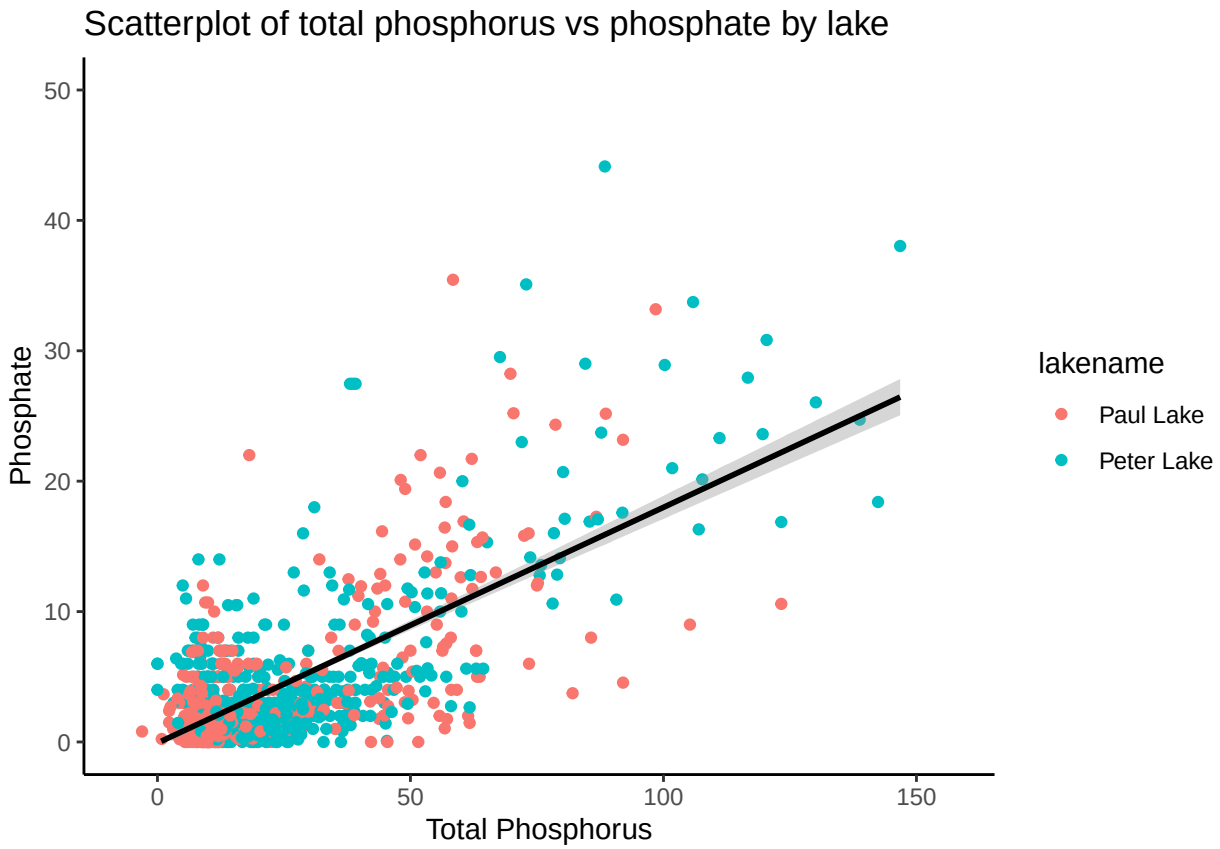
```
#4
plot1 <- ggplot(NCP, aes(x = tp_ug, y = po4, color = lakename))+
  geom_point()+
  geom_smooth(method = "lm", colour="black")+
  ylim(0, 50) +
  ggtitle("Scatterplot of total phosphorus vs phosphate by lake") +
  xlab("Total Phosphorus")+
  ylab("Phosphate")
print(plot1)
```

```
## `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).
```

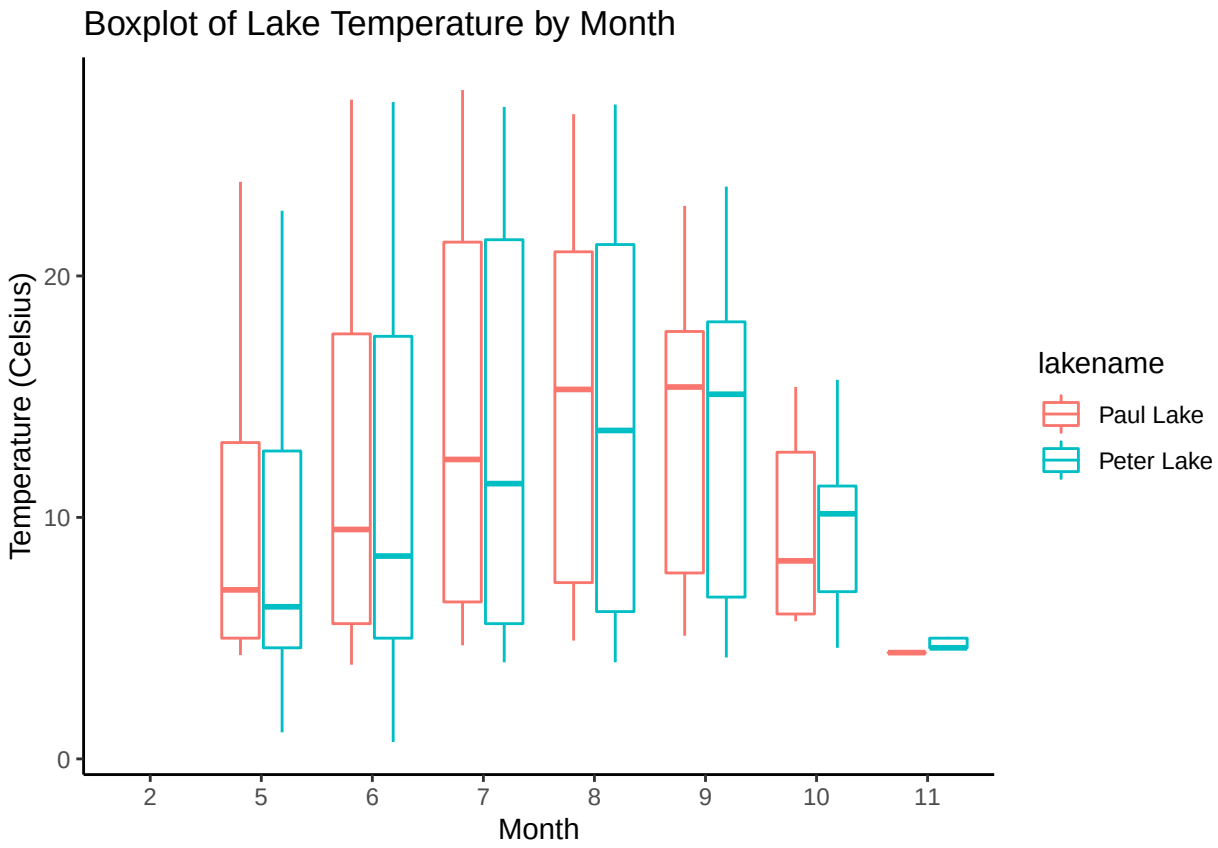
```
## Warning: Removed 2 rows containing missing values (geom_smooth).
```



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.

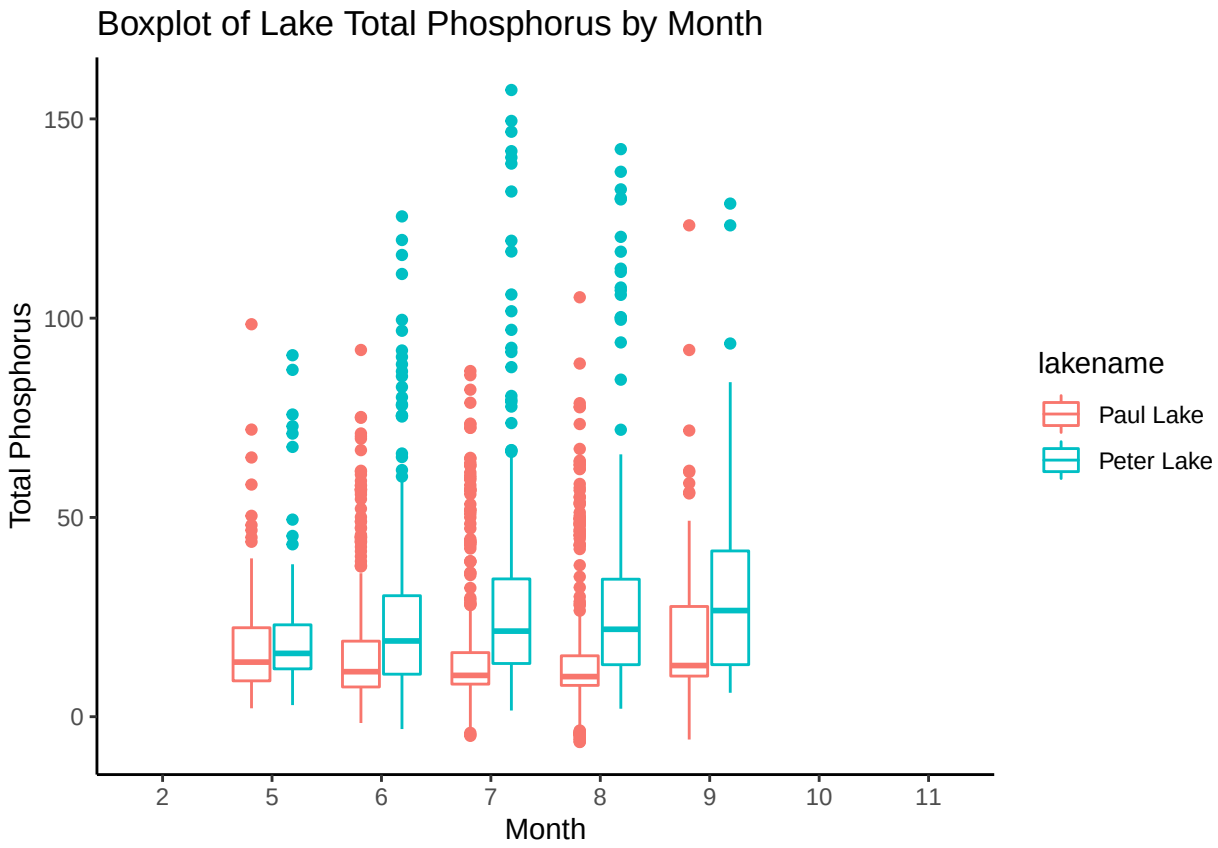
```
#5
NCP$month <- as.factor(NCP$month)

plot2a <- ggplot(NCP, aes(x = month, y = temperature_C, color = lakename))+
  geom_boxplot() +
  ggtitle("Boxplot of Lake Temperature by Month") +
  xlab("Month")+
  ylab("Temperature (Celsius)")
print(plot2a)
```

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



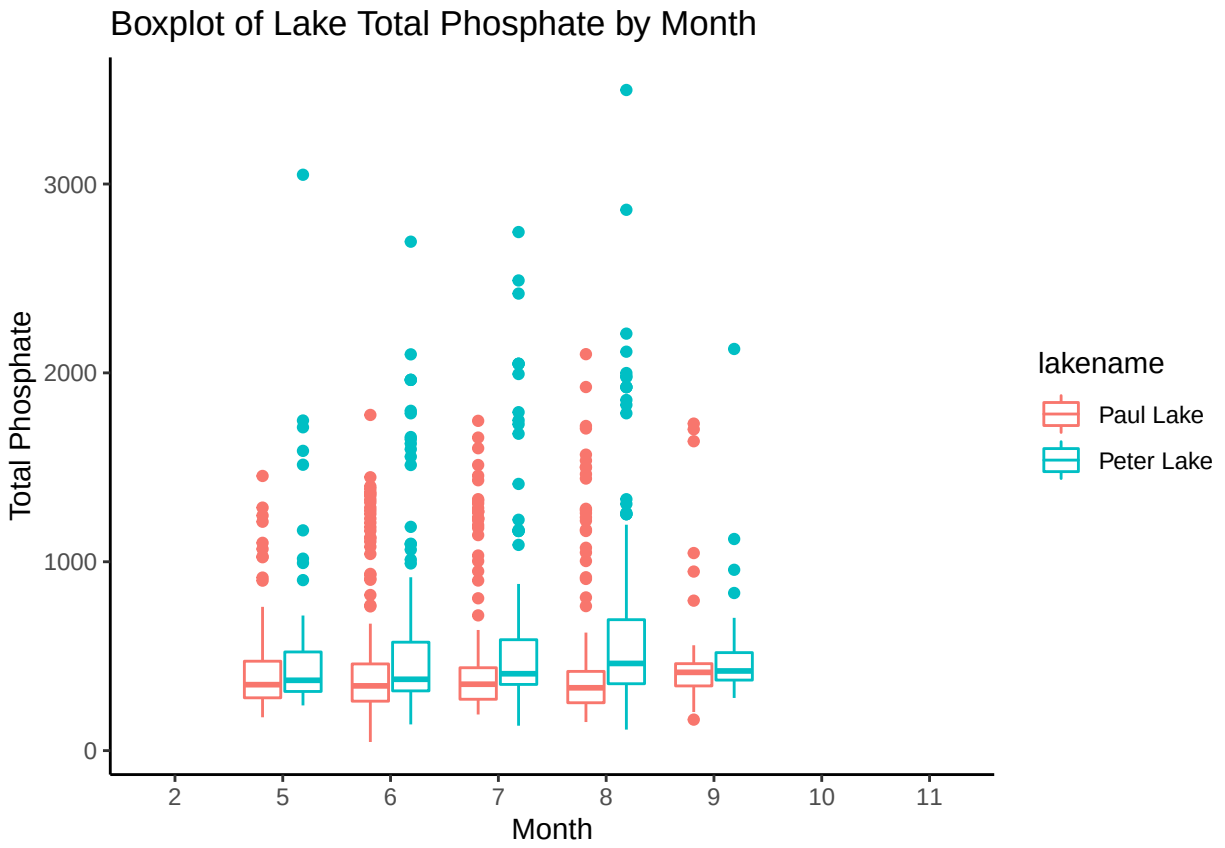

```
plot2b <- ggplot(NCP, aes(x = month, y = tp_ug, color = lakename))+
  geom_boxplot() +
  ggtitle("Boxplot of Lake Total Phosphorus by Month") +
  xlab("Month")+
  ylab("Total Phosphorus")
print(plot2b)
```

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




```
plot2c <- ggplot(NCP, aes(x = month, y = tn_ug, color = lakename))+
  geom_boxplot() +
  ggtitle("Boxplot of Lake Total Phosphate by Month") +
  xlab("Month")+
  ylab("Total Phosphate")
print(plot2c)
```

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




```
plot2 <- plot_grid(plot2a + theme(legend.position = "none"),
                   plot2b + theme(legend.position = "none"),
                   plot2c + theme(legend.position = "none"),
                   labels = c("a", "b", "c"),
                   align = 'h', axis = 'l', nrow = 1)
```

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

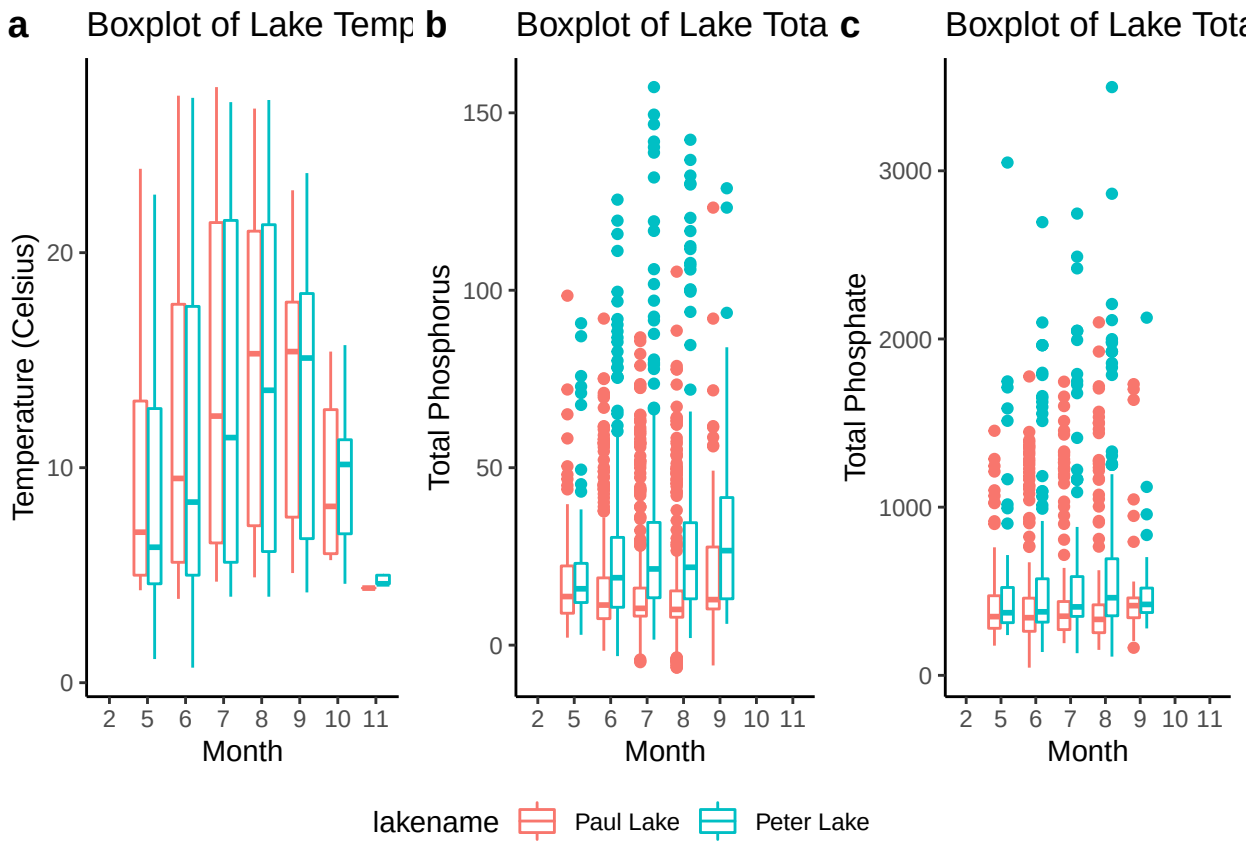
```
## Warning: Removed 20729 rows containing non-finite values (stat_boxplot).
```

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

```
legend <- get_legend(
  plot2a +
    guides(color = guide_legend(nrow = 1)) +
    theme(legend.position = "bottom"))
```

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

```
plot_grid(plot2, legend, ncol = 1, rel_heights = c(1,0.1))
```



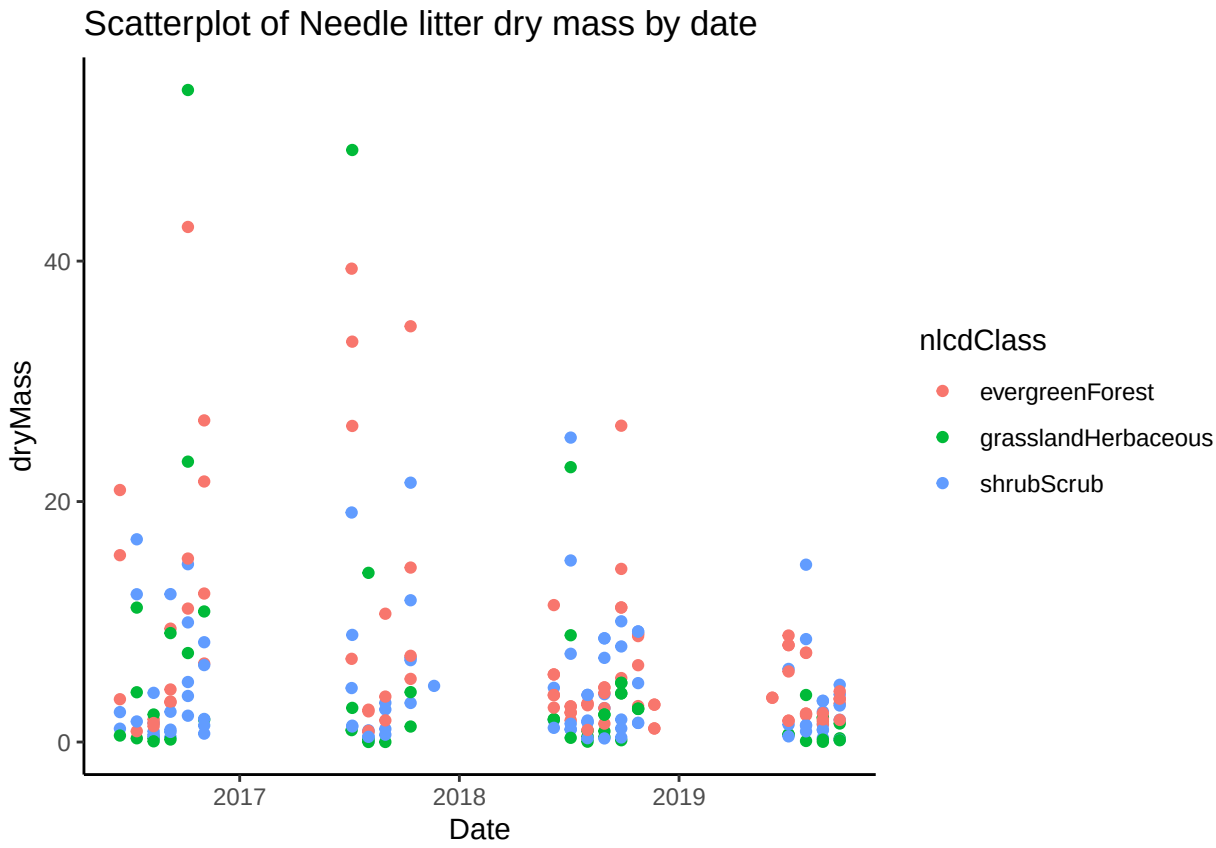
Question: What do you observe about the variables of interest over seasons and between lakes?

Answer: Compared to the Paul lake, Peter Lake usually have a higher tested amount of chemicals regardless of the season. The tested amount chemicals tend to increase as temperature rises and reach the peaks during the summer months, July and August.

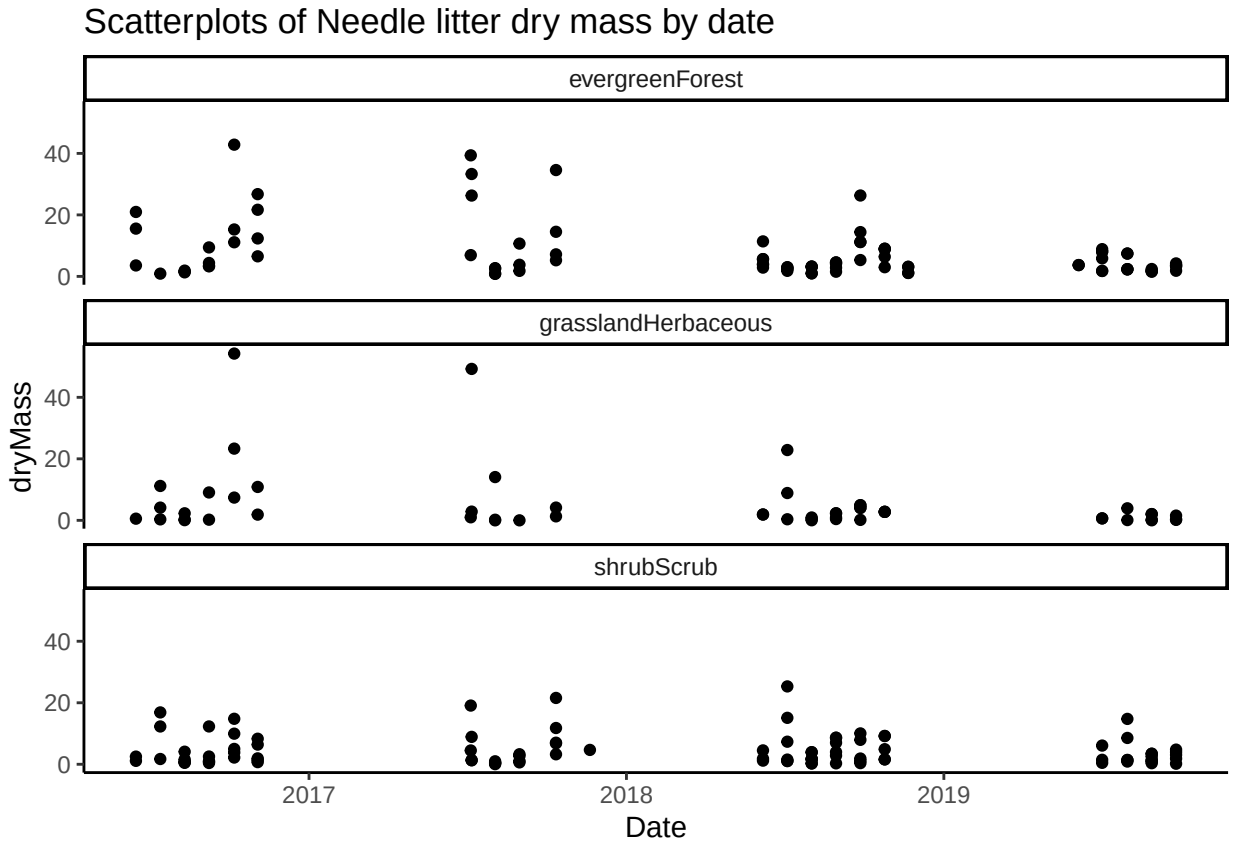
6. [Niwot Ridge] Plot a subset of the litter dataset by displaying only the "Needles" functional group. Plot the dry mass of needle litter by date and separate by NLCD class with a color aesthetic. (no need to adjust the name of each land use)

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

```
#6
plot3 <-
ggplot(subset(NIL, functionalGroup == 'Needles'),
       aes(x = collectDate, y = dryMass, color = nlcdClass))+
  geom_point()+
  ggtitle("Scatterplot of Needle litter dry mass by date") +
  xlab("Date")+
  ylab("dryMass")
print(plot3)
```




```
#7
plot4 <-
  ggplot(subset(NIL, functionalGroup == 'Needles'),
         aes(x = collectDate, y = dryMass))+
  geom_point() +
  facet_wrap(vars(nlcdClass), nrow = 3)+
  ggtitle("Scatterplots of Needle litter dry mass by date") +
  xlab("Date")+
  ylab("dryMass")
print(plot4)
```



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

Answer: According to the bar chart, boxplot, and violin plot. None of the three could display the value of each single observation data as dot plot does. I think plot4 is more effective because it is easier to read, both seperately and together, and it can easily present the change within each of the classes.