Assignment 5: Data Visualization

Vicky Jia

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., "Salk_A05_DataVisualization.Rmd") prior to submission.

The completed exercise is due on Tuesday, February 11 at 1:00 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 (tidy and gathered) 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
getwd()
```

```
## [1] "/Users/Victoria/Environmental_Data_Analytics_2020"
```

```
library(tidyverse)
library(cowplot)
PPchem <- read.csv("./Data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv")
PPgather <- read.csv('./Data/Processed/NTL-LTER_Lake_Nutrients_PeterPaulGathered_Processed.csv')
litter <- read.csv('./Data/Processed/NEON_NIWO_Litter_mass_trap_Processed.csv')

#2
PPchem$sampledate <- as.Date(PPchem$sampledate,format = '%Y-%m-%d')
PPgather$sampledate <- as.Date(PPgather$sampledate,format = '%Y-%m-%d')
litter$collectDate <- as.Date(litter$collectDate,format = '%Y-%m-%d')
```

Define your theme

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

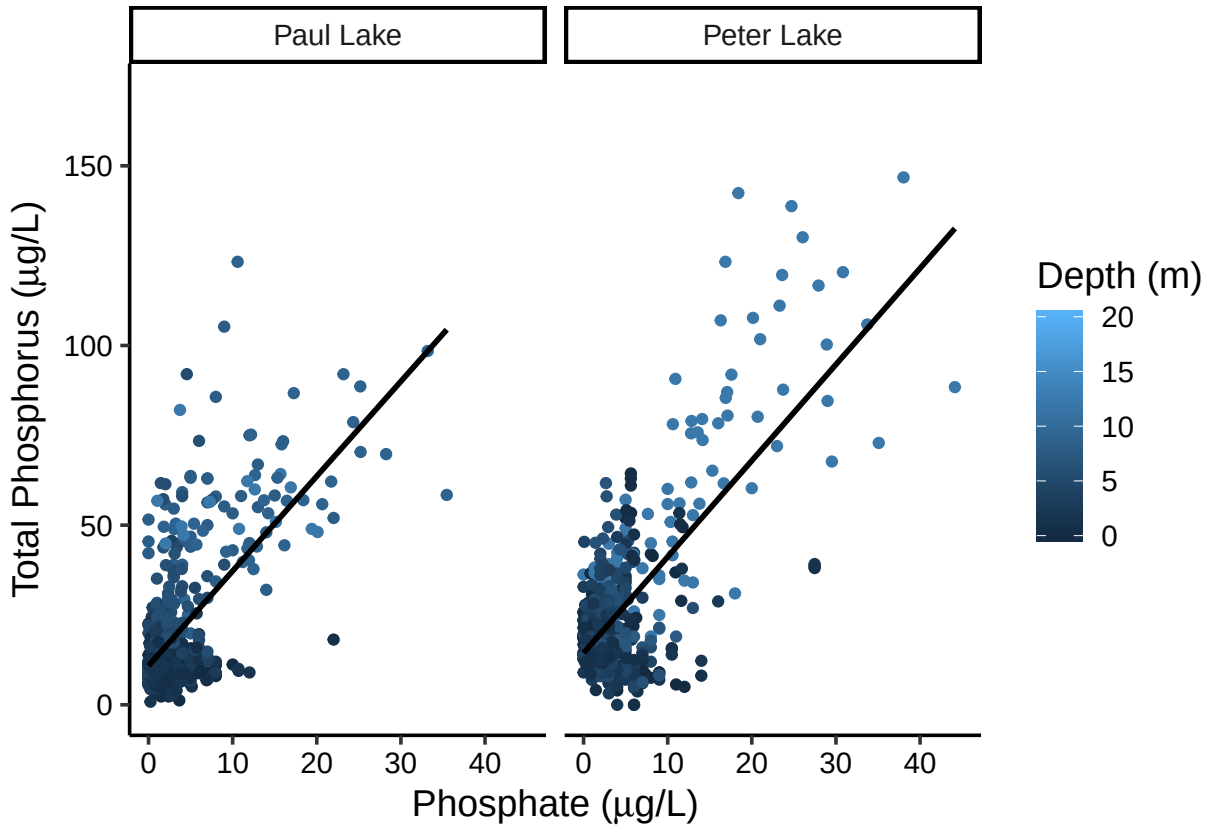
```
mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"))

theme_set(mytheme)
```

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 by phosphate, 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.

```
TPpo <-
  ggplot(PPchem, aes(x = po4, y = tp_ug, color = depth)) +
  geom_point()+
  xlim(0, 45) +
  ylim(0, 170)+
  geom_smooth(method = lm, color = 'black', se = FALSE) +
  labs(y = expression(paste("Total Phosphorus (", mu, "g/L)")),
       x= expression(paste("Phosphate (",mu, "g/L)")),
       color = "Depth (m)", shape = "")+
  facet_grid(~lakename)
print(TPpo)
```



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.

```
tempplot <- ggplot(PPchem, aes(x = as.factor(month), y = temperature_C, color = lakename))+
  geom_boxplot()+
  labs(y = 'Temperature (C)',
       x= 'Month')

tnplot <- ggplot(PPchem, aes(x = as.factor(month), y = tn_ug, color = lakename))+
  geom_boxplot()+
  labs(y = expression(paste("TN (", mu, "g/L)")),
```

```
       x= 'Month')

tpplot <-  ggplot(PPchem, aes(x = as.factor(month), y = tp_ug, color = lakename))+
  geom_boxplot()+
  labs(y = expression(paste("TP (", mu, "g/L)")),
       x= 'Month')

comb <- plot_grid(tempplot +  theme(legend.position = "none"),
          tnplot +  theme(legend.position = "none"),
          tpplot +  theme(legend.position = "none"),
          nrow = 1, align = 'vh', hjust = -1)
```

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

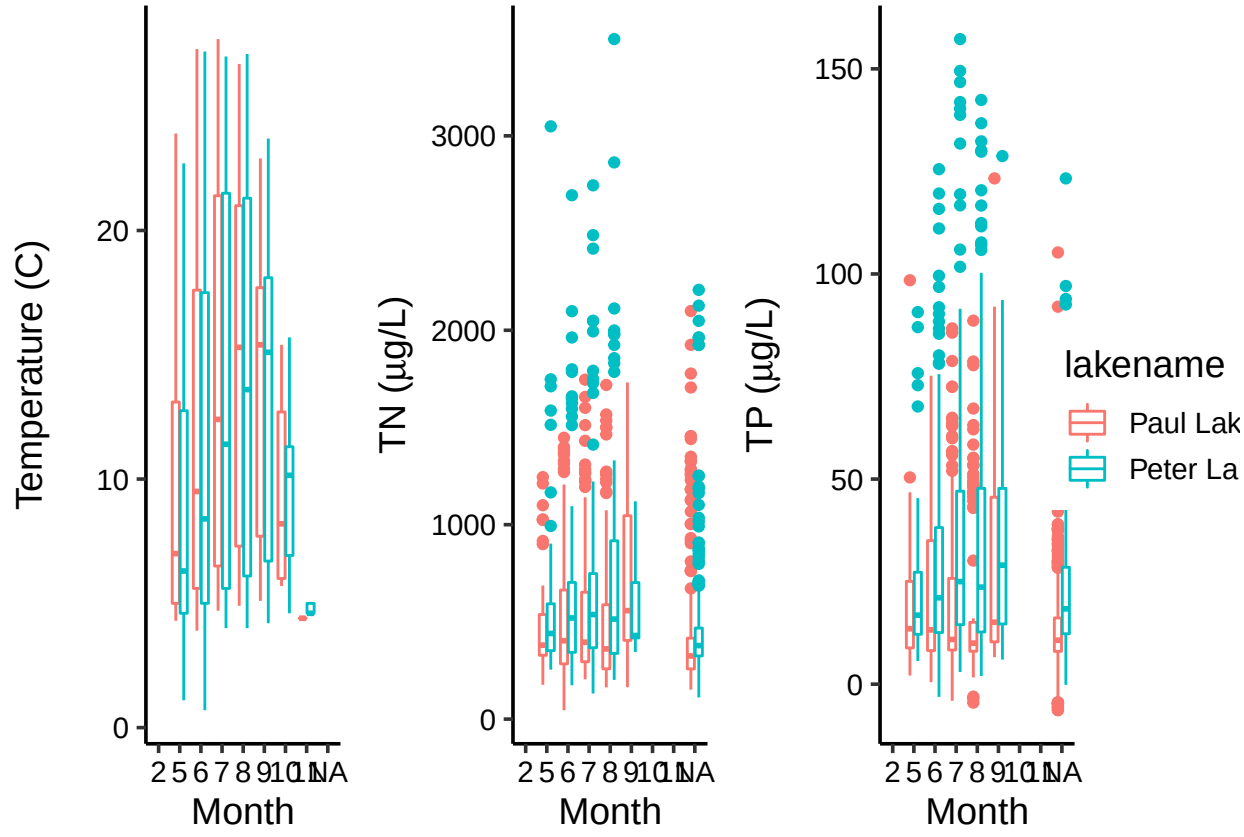
```
## Warning: Removed 21583 rows containing non-finite values (stat_boxplot).
```

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

```
legend <- get_legend(
  tnplot + theme(legend.position = "right"))
```

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

```
plot_grid(comb, legend, rel_widths = c(3, 0.4))
```



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

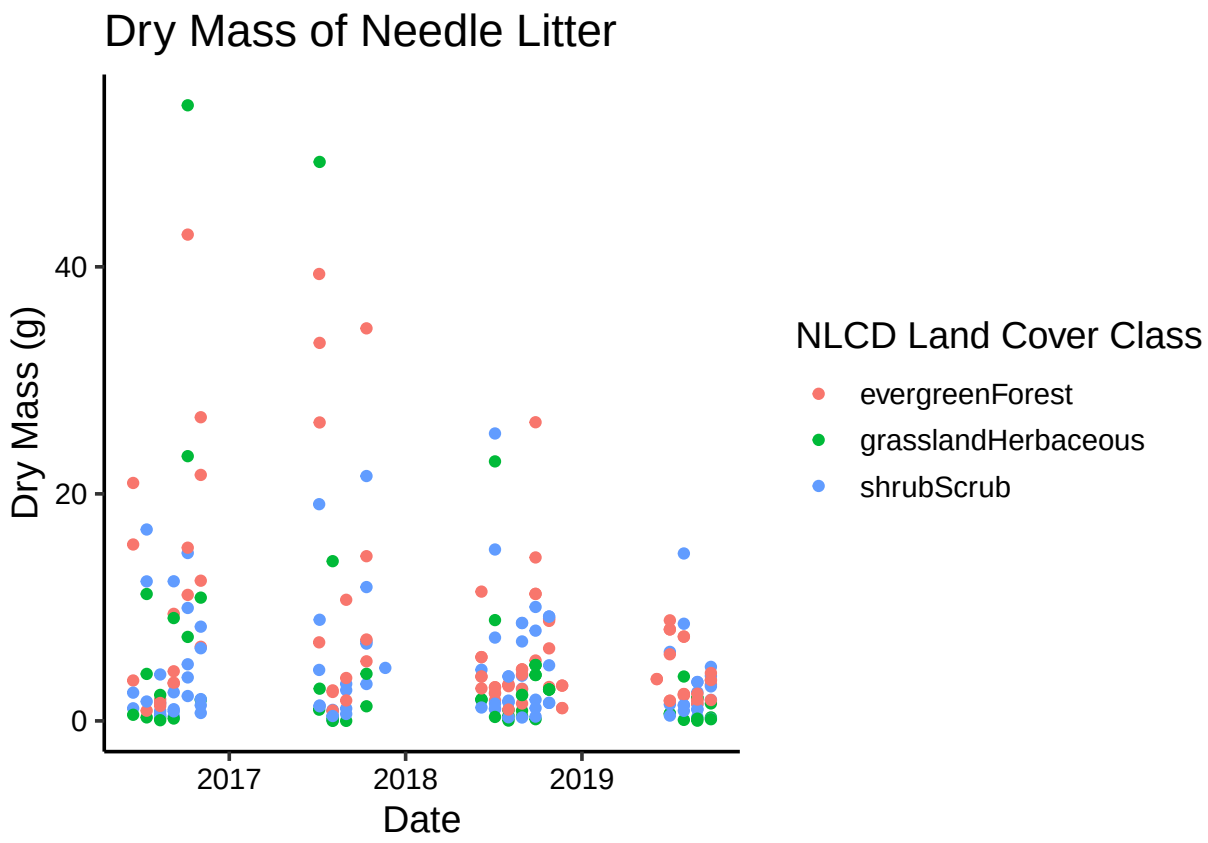
Answer: Most data are available in summer (from May to October) and fewer are available in Febuary and November.Paul Lake has fewer outliers than Peter Lake for nutrients, while there is

no outlier for temperature data. Also Peter Lake has a wider interquartile (IQR) range than 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)

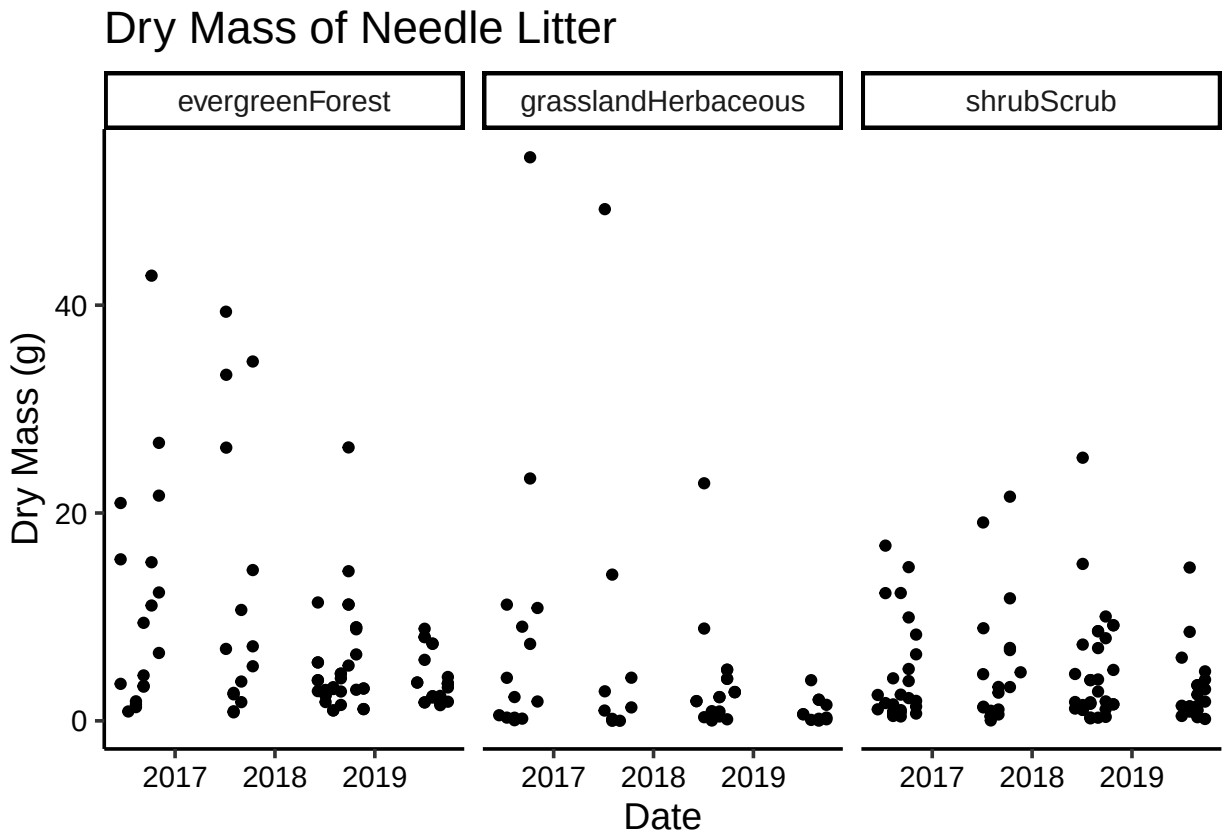
7. [Niwot Ridge] Now, plot the same plot but with NLCD classes separated into three facets rather than separated by color.

```
litterNeedles <-
  ggplot(subset(litter, functionalGroup == "Needles"),
         aes(x = collectDate, y = dryMass, color = nlcdClass))+
  geom_point()+
  labs(y = 'Dry Mass (g)',
       x= 'Date',
       title = 'Dry Mass of Needle Litter',
       color = "NLCD Land Cover Class", shape = "")

print(litterNeedles)
```




```
litterNeedlesNLCD <-
  ggplot(subset(litter, functionalGroup == "Needles"),
         aes(x = collectDate, y = dryMass))+
  geom_point()+
  labs(y = 'Dry Mass (g)',
       x= 'Date',
       title = 'Dry Mass of Needle Litter')+
```

```
  facet_grid(~nlcdClass)

print(litterNeedlesNLCD)
```



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

Answer: In this question, I think the facet plot is more effective, because it clearly displays the distribution of the dry mass of needle litter each year, and also allows me to compare each land cover both parallelly and vertically - that says, I could use this fact plot to both compare the same land cover at different times and compare different land covers at the same time. Admittedly, we only have three land cover variables here so the color theme can also illustrate differences over dry mass over the time. However, if we have more than three varibles, the color differentiated graph will be hard to read.