Assignment 5: Data Visualization

Michael Gaffney

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
#load libraries
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)
library(viridis)
```

```
## Loading required package: viridisLite
```

```
library(RColorBrewer)
library(colormap)
setwd("/Users/michaelgaffney/Documents/Duke University/Nicholas School of the Environment/05 Spring 202
```

```r
#load data
PeterPaul <- read.csv("./data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv", str
Litter <- read.csv("./data/Processed/NEON_NIWO_Litter_mass_trap_Processed.csv", stringsAsFactors = TRUE
#2 Check dates for both datasets
class(PeterPaul$sampledate)
```

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

```r
class(Litter$sampledate)
```

```
## [1] "NULL"
```

```r
#both read as factors; transform both to dates
PeterPaul$sampledate <- as.Date(PeterPaul$sampledate, format = "%Y-%m-%d")
Litter$sampledate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
```

Define your theme

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

```r
#3 create dark theme variable.
mytheme <- theme_dark(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
#set the theme globally
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 (`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()`).

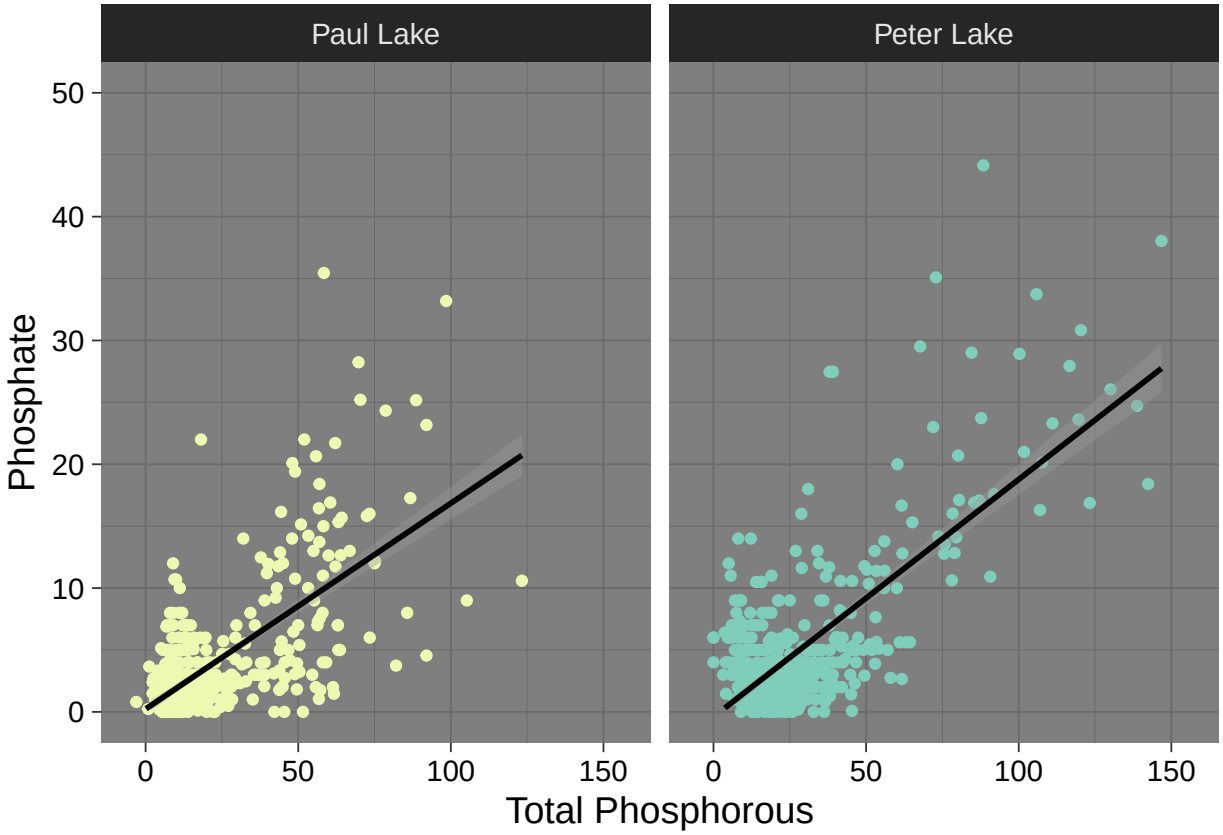
```r
#4
ggplot(PeterPaul, aes(x = tp_ug, y = po4, color = lakename)) +
  #create two side by side graphs with Peter and Paul
  facet_wrap("lakename") +
  #remove legend and add points
  geom_point(show.legend = FALSE) +
  geom_smooth(method = lm, color = "black") +
  xlab("Total Phosphorous") +
  ylab("Phosphate") +
  ylim (0,50) +
  scale_color_brewer(palette = "YlGnBu")
```

```
## `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 4 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 Create three plots
PeterPaul_Temp <- ggplot(PeterPaul, aes(x = as.factor(month), y = temperature_C, color = lakename)) +
  geom_boxplot() +
  xlab("Month") + ylab("Temperature (C)") +
  scale_color_manual(values = c("hotpink", "lightgreen")) +
  labs(color = "Lakes")

PeterPaul_TP <- ggplot(PeterPaul, aes(x = as.factor(month), y = tp_ug, color = lakename)) +
  geom_boxplot(show.legend = FALSE) +
  xlab("Month") + ylab("TP") +
  scale_color_manual(values = c("hotpink", "lightgreen"))

PeterPaul_TN <- ggplot(PeterPaul, aes(x = as.factor(month), y = tn_ug, color = lakename)) +
  geom_boxplot(show.legend = FALSE) +
  xlab("Month") + ylab("TN") +
  scale_color_manual(values = c("hotpink", "lightgreen"))

#grab the legend from the first plot
my.legend <- get_legend(PeterPaul_Temp)
```

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

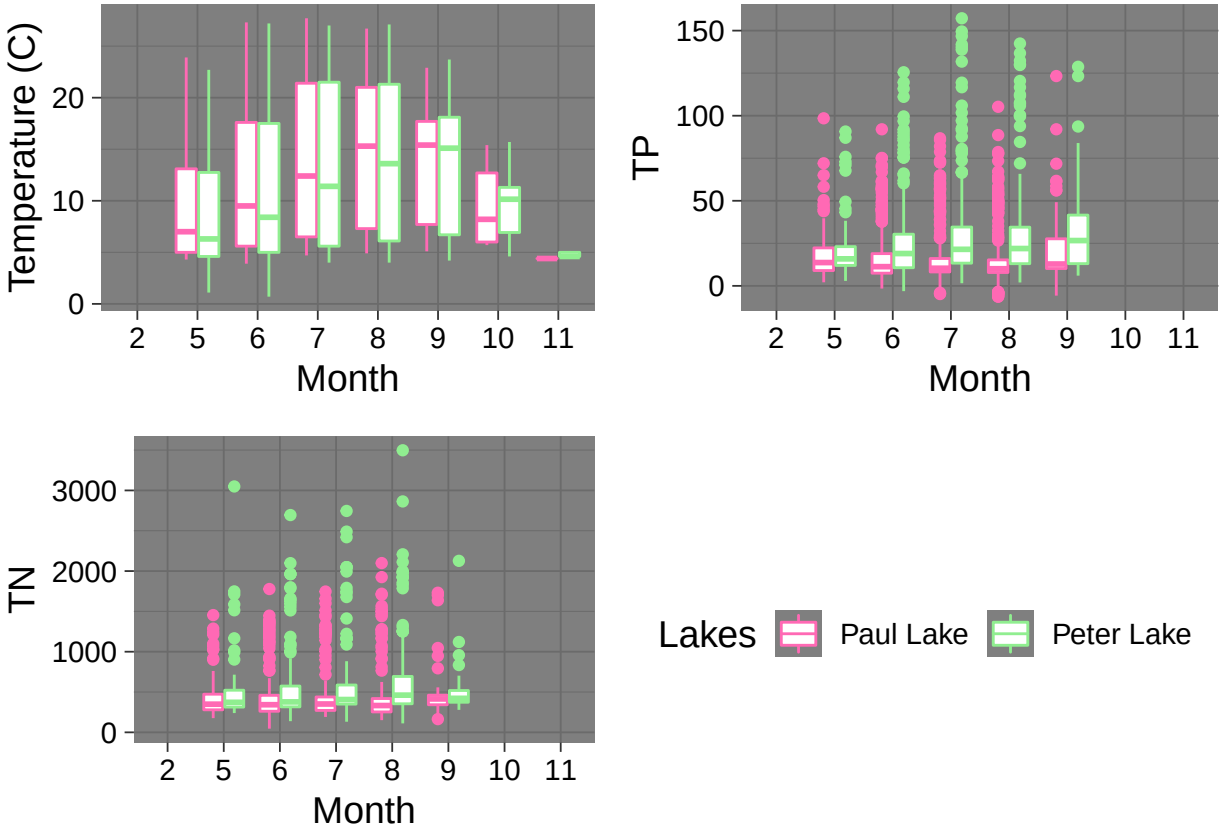
```
#now remove legend and replot without the legend
PeterPaul_Temp <- PeterPaul_Temp + theme(legend.position = "none")
#plot the grid
```

```
plot_grid(PeterPaul_Temp, PeterPaul_TP, PeterPaul_TN, my.legend,
          ncol = 2, align = "vh", axis = "b")
```

```
## 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).
```



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

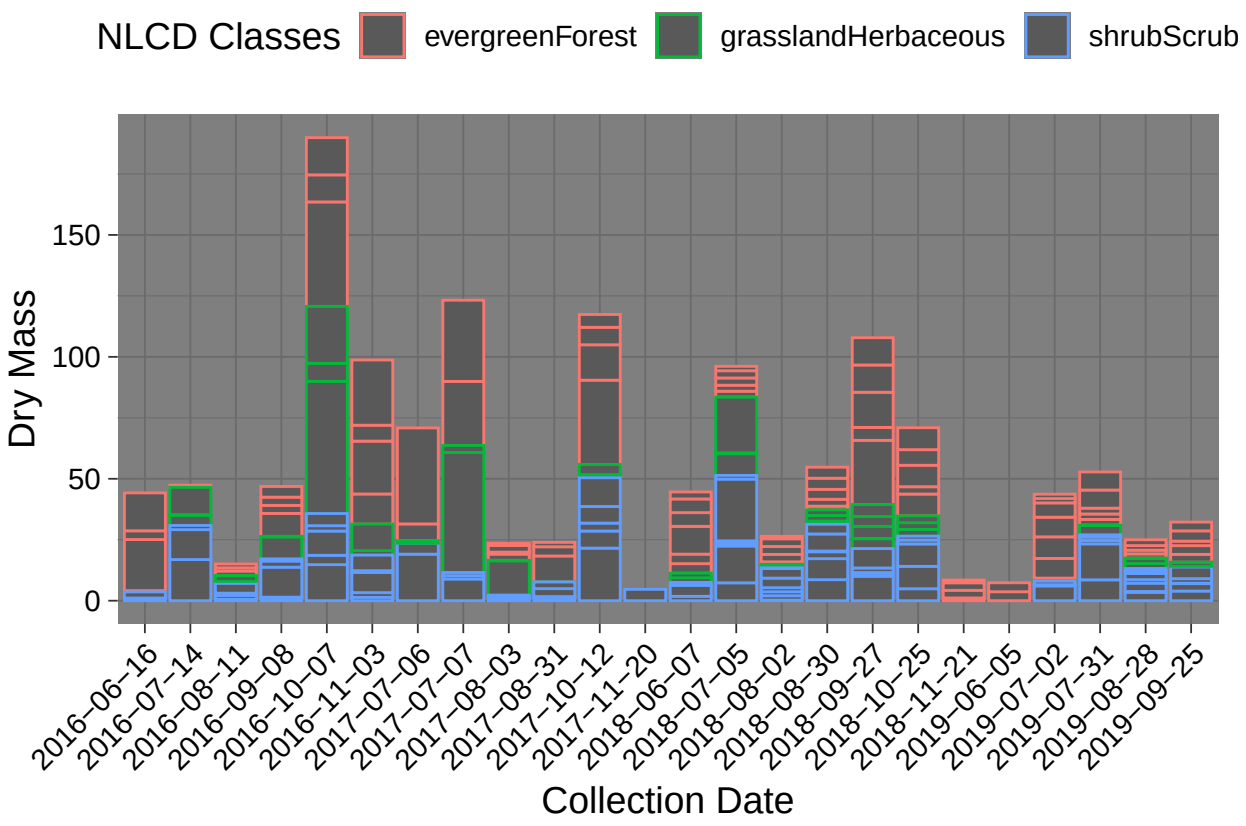
> Answer: The temperature for both lakes increases over the summer months, as we would expect. Temperature seems realtively consistent accross the two lakes, with the most noticeable difference happening in October (Peter Lake is much warmer). With TP, there are more noticeable differences. Peter Lake has noticeablely higher values throughout most of the year, except for May. There is some variability for both lakes across the year, although the trends are not consistent between lakes: Peter Lake seems to increase in TP value over the year, while Paul lake appears to decrease and increase over the year. For TN, values for both lakes seem to remain stable accross the year, with a loose trend for Peter Lake to increase over the year. Peter Lake tends to have higher median values 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)

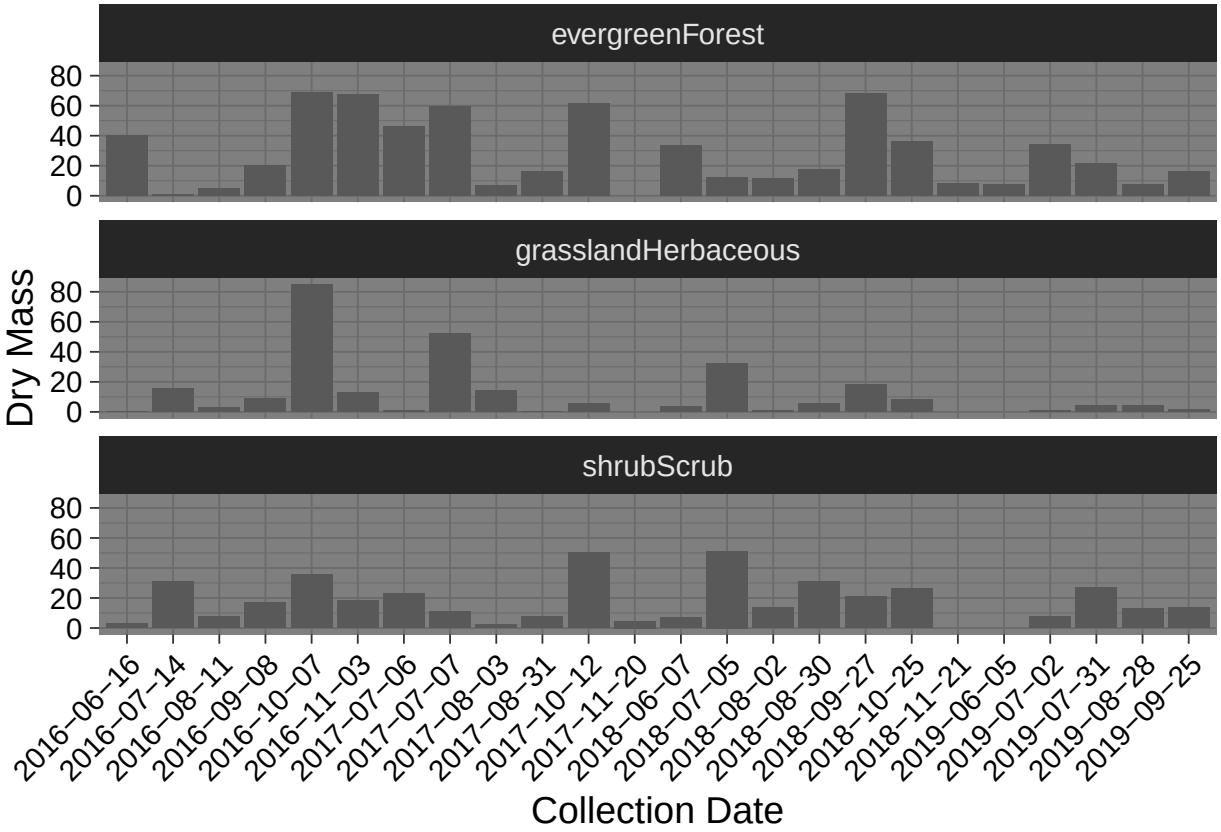
7. [Niwot Ridge] Now, plot the same plot but with NLCD classes separated into three facets rather than separated by color.

```
#6plot with dry mass by date, with
ggplot(filter(Litter, functionalGroup == "Needles")) +
```

```
  geom_col(aes(x = collectDate, y = dryMass, color = nlcdClass)) +
  #sets the text on the x axis at an angle so it is readable.
  theme(axis.text.x = element_text(angle = 45,  hjust = 1)) +
  #change label names
  xlab("Collection Date") + ylab("Dry Mass") +
  #Change the legend title; this code makes absolutely no sense
  labs(color = "NLCD Classes")
```




```
#7 Seconf plot, but with NLCD classes separated out in different plots
ggplot(filter(Litter, functionalGroup == "Needles")) +
  geom_col(aes(x = collectDate, y = dryMass)) +
  facet_wrap(vars(nlcdClass), nrow = 3) +
  #sets the text on the x axis at an angle so it is readable.
  theme(axis.text.x = element_text(angle = 45,  hjust = 1)) +
  xlab("Collection Date") + ylab("Dry Mass") +
  #Change the legend title; this code makes absolutely no sense
  labs(color = "NLCD CLasses")
```



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

Answer: If we are interested in dry mass for different land covers, the second plot with facets is substantially more effective. By breaking the data up by landcover, it becomes easier to see the patterns for each type of cover over time. With the first graph, the nuances of change by cover type over the collection period are largely lost. But, if you were more interested in those collected values, and only had a passing interest in landcover, then it is possible that the first graph would work better.