Assignment 5: Data Visualization

Ina Liao

Fall 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. Read in 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 in the Processed_KEY folder) and the processed data file for the Niwot Ridge litter dataset (use the `NEON_NIWO_Litter_mass_trap_Processed.csv` version, again from the Processed_KEY folder).

```
#install.packages("tidyverse")
#install.packages("lubrudate")
#install.packages("here")
#install.packages("ggridges")
#install.packages("viridis")
#install.packages("RColorBrewer")
#install.packages("colormap")
#install.packages("ggthemes")
#install.packages("cowplot")
```

```
library(tidyverse)
library(lubridate)
library(here)
library(ggridges)
library(viridis)
library(RColorBrewer)
library(colormap)
library(ggthemes)
library(cowplot)
```

```
processed_data="/Users/inaliao/Desktop/EDE_Fall2023/Data/Processed_Key"

#1
df_NTL<-read.csv(here(processed_data,"NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv"))

#2
df_Niwot<-read.csv(here(processed_data,"NEON_NIWO_Litter_mass_trap_Processed.csv"))
```

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_plot_theme<- theme_calc()+
  theme(
    #plot background
    plot.background = element_rect(color="gray"),

    #plot title
    plot.title=element_text(color="black",hjust=0.5,vjust=1),

    #axis labels
    axis.text=element_text(color="black"),

    #gridlines
    panel.grid.major=element_line("white"),
    axis.ticks=element_line(color="white"),

    #legend
    legend.key=element_rect(color="white"),
    legend.background = element_rect(color="white"),
    legend.position="right"
      )
```

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()`).

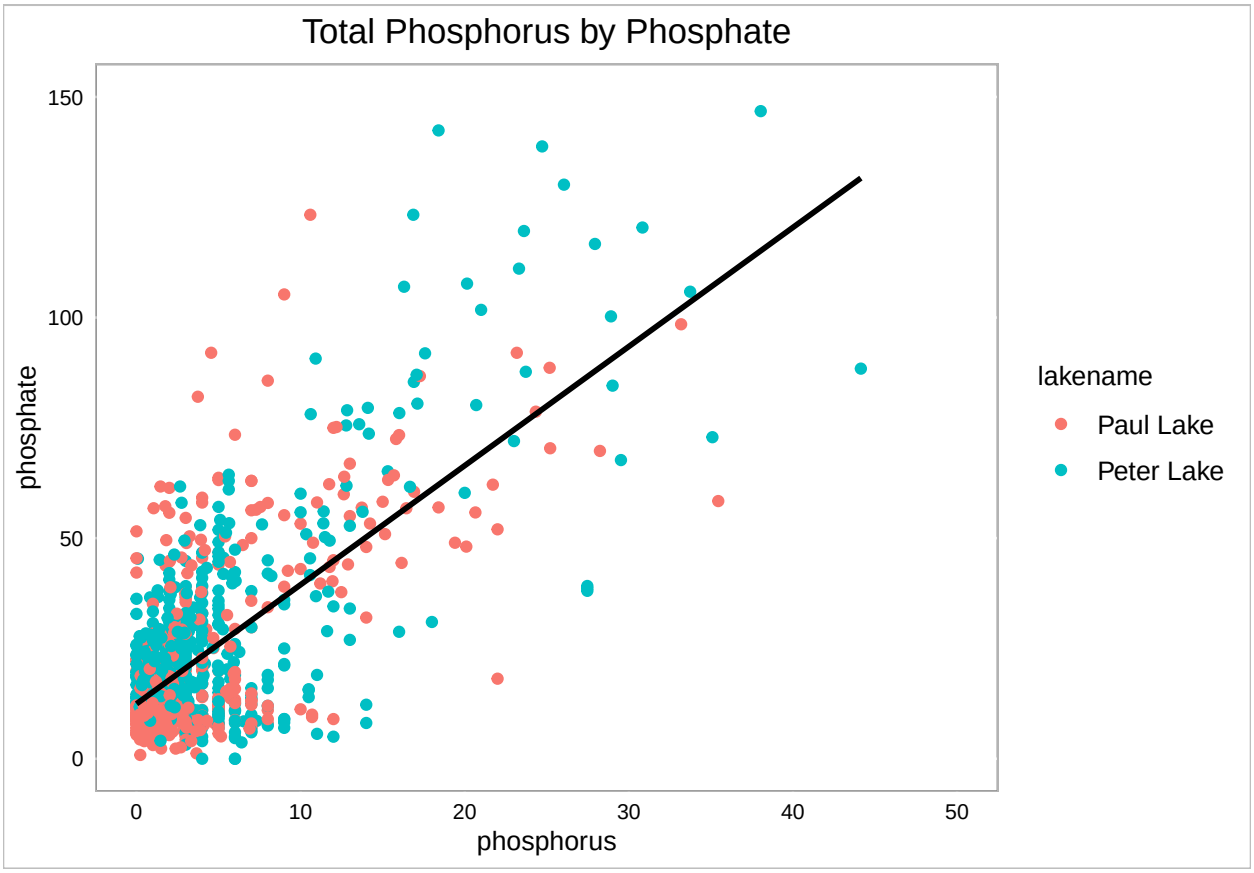
```
#4
ggplot(subset(df_NTL,lakename=="Paul Lake" | lakename=="Peter Lake"))+
  geom_point(aes(x=po4,y=tp_ug,color=lakename))+
  xlim(0,50)+ylim(0,150)+ #add limits
  ggtitle("Total Phosphorus by Phosphate")+
  labs(x="phosphorus",y="phosphate")+
```

```
  geom_smooth(aes(x=po4,y=tp_ug),method = "lm",se=FALSE,color="black")+ #add a trend line
  my_plot_theme #defined theme
```

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

```
## Warning: Removed 21948 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 21948 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: * Recall the discussion on factors in the previous section as it may be helpful here. * R has a built-in variable called `month.abb` that returns a list of months;see https://r-lang.com/month-abb-in-r-with-example

```
#5

#boxplot of temperature
plot_temp<-ggplot(df_NTL)+
  geom_boxplot(aes(x=factor(month,levels=1:12,labels=month.abb),y=temperature_C,color=lakename))+
  scale_x_discrete(name="month",drop=FALSE)+
  labs(x="month",y="temperature (Celcius)")+
  my_plot_theme+
  theme(axis.title.x = element_text(size = 8),
        axis.title.y = element_text(size = 8),
        axis.text.x = element_text(size = 6),
        axis.text.y = element_text(size = 6),
```

```
         legend.title = element_text(size = 8),
        legend.text = element_text(size = 8))

#boxplot of TP
plot_TP<-ggplot(df_NTL)+
  geom_boxplot(aes(x=factor(month,levels=1:12,labels=month.abb),y=tp_ug,color=lakename))+
  scale_x_discrete(name="month",drop=FALSE)+
  labs(x="month",y="phosphorus (mg/dL)")+
  my_plot_theme+
  theme(axis.title.x = element_text(size = 8),
        axis.title.y = element_text(size = 8),
        axis.text.x = element_text(size = 6),
        axis.text.y = element_text(size = 6),
        legend.title = element_text(size = 8),
        legend.text = element_text(size = 8))


#boxplot of TN
plot_TN<-ggplot(df_NTL)+
  geom_boxplot(aes(x=factor(month,levels=1:12,labels=month.abb),y=tn_ug,color=lakename))+
  scale_x_discrete(name="month",drop=FALSE)+
  labs(x="month",y="phosphate (mg/dL)")+
  my_plot_theme+
  theme(axis.title.x = element_text(size = 8),
        axis.title.y = element_text(size = 8),
        axis.text.x = element_text(size = 6),
        axis.text.y = element_text(size = 6),
        legend.title = element_text(size = 8),
        legend.text = element_text(size = 8)
        )
```

```
#to ensure there is only one legend in the combined plot
plot_TP<-plot_TP+theme(legend.position="none")
plot_TN<-plot_TN+theme(legend.position="none")

# use cowplot to combine three boxplots
plot_combine<-plot_grid(plot_temp,plot_TP,plot_TN,
                        ncol=1,aligh="hy",
                        labels=c("Temerature","Phosphorus","Phosphate"),
                        label_size=10)
```

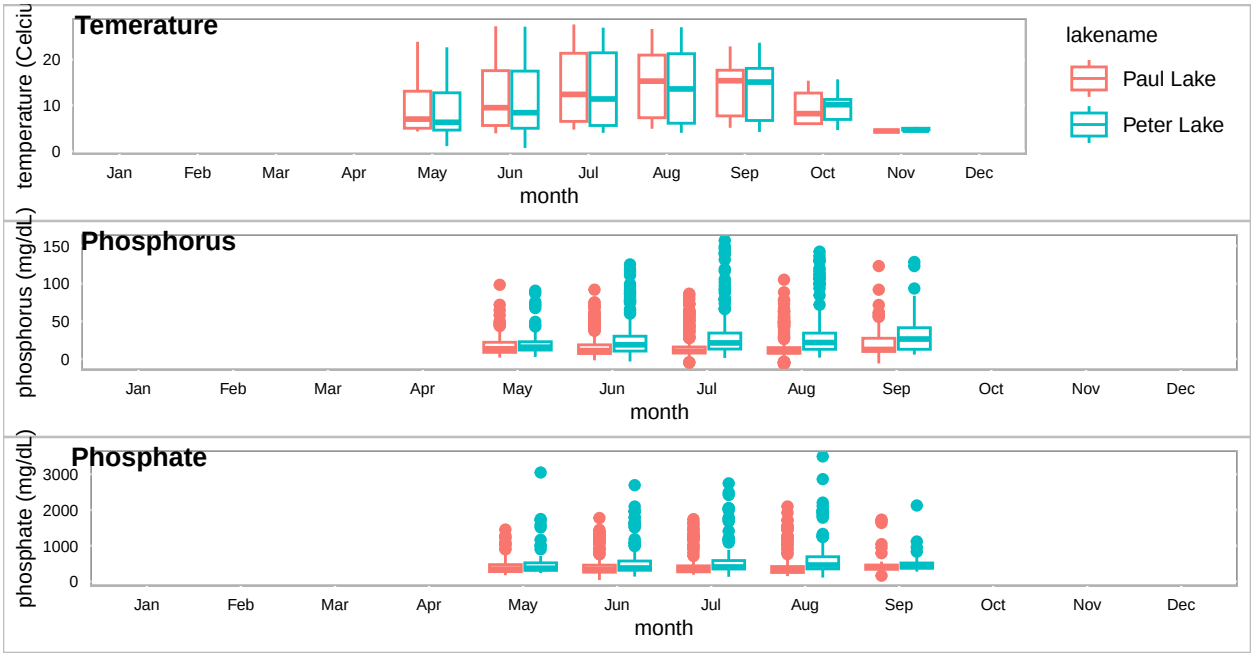
```
## Warning: Removed 3566 rows containing non-finite values (`stat_boxplot()`).

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

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

## Warning in as_grob.default(plot): Cannot convert object of class character into
## a grob.
```

```
plot_combine
```



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

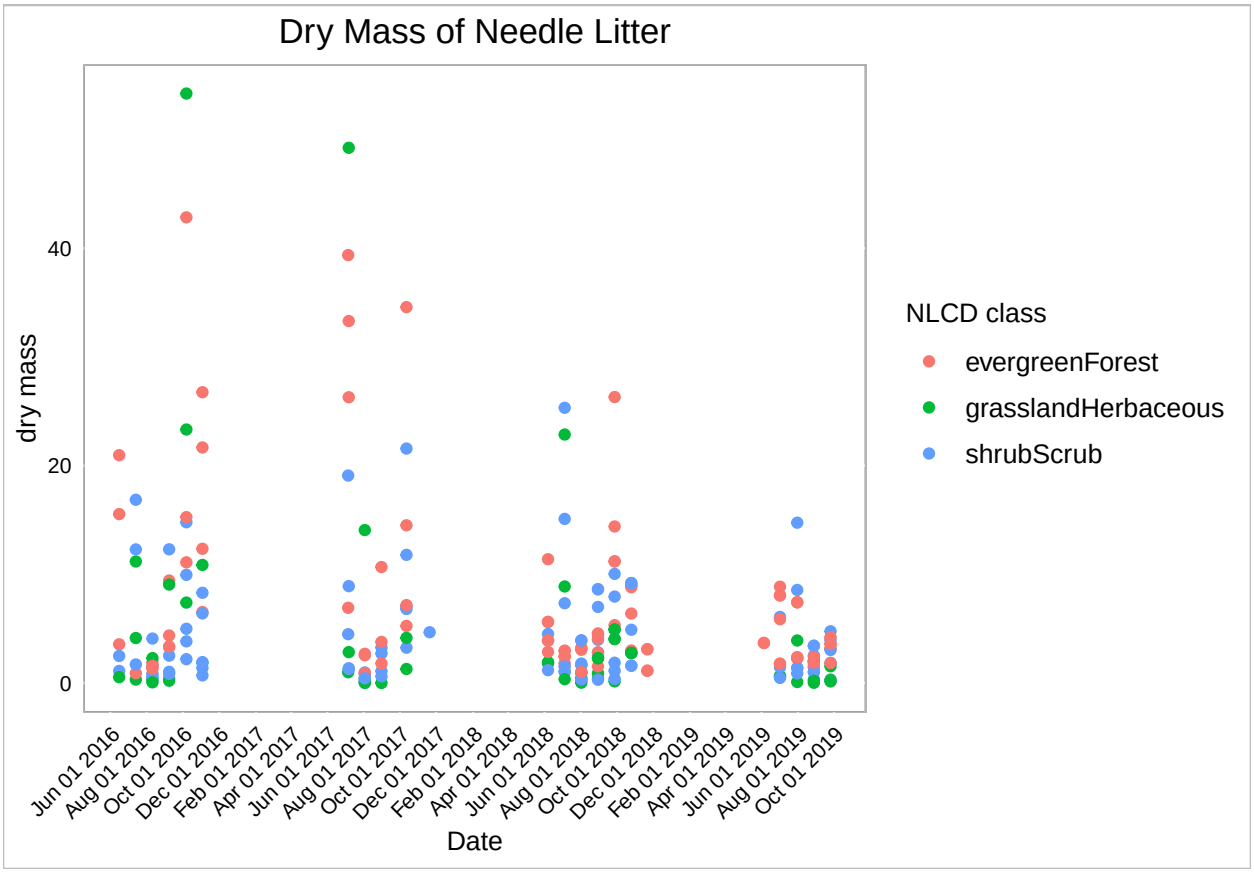
Answer: The temperature distribution in Paul Lake and Peter Lake is relatively similar. Both Paul Lake and Peter Lake have the highest average temperatures in September. In addition, during July and August, the temperature has the highest range, which might result from differences in temperature between the surface and deeper layers of the lake. Both phosphorus and phosphate in Peter Lake are averagely higher than in Paul Lake. In both Paul Lake and Peter Lake in June, July, and August, phosphorus and phosphate show the most outliers.

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)

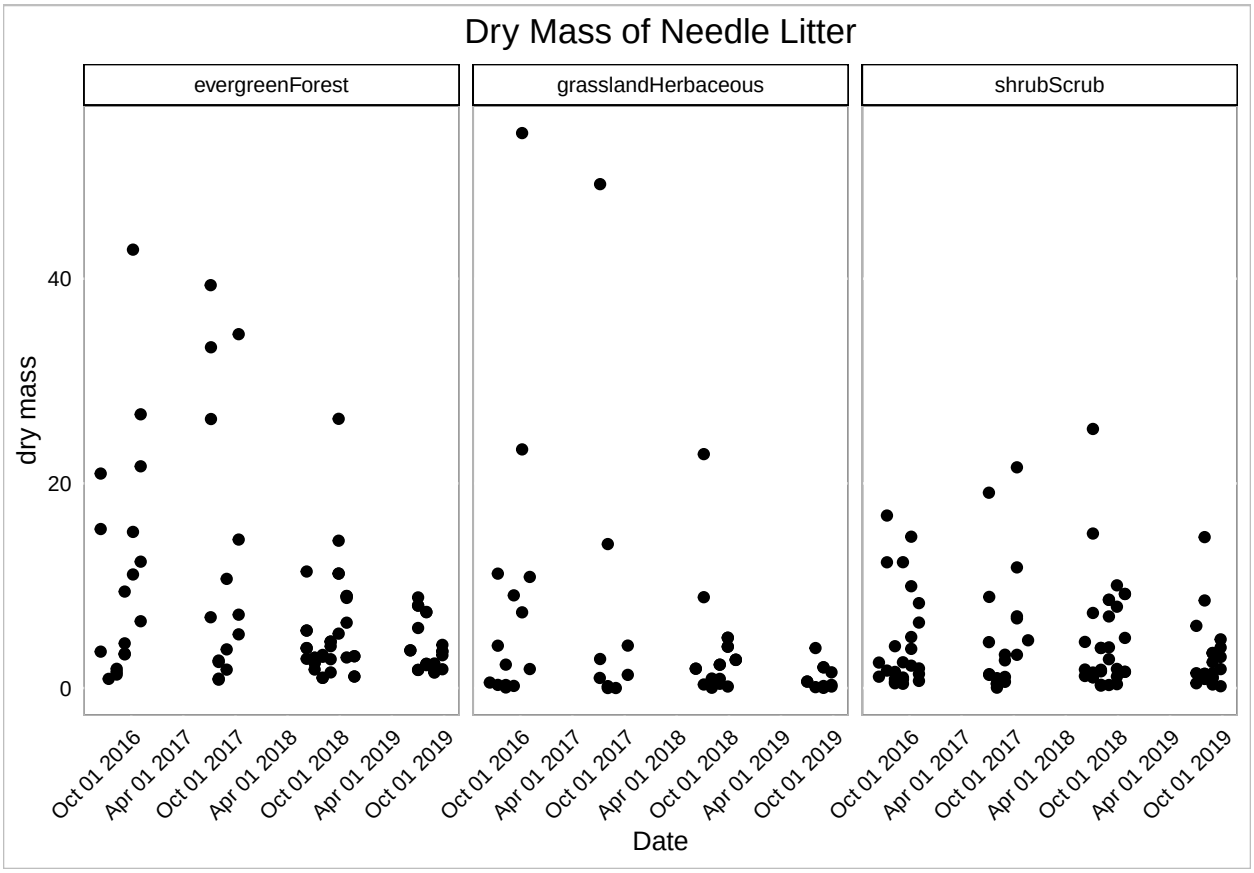
7. [Niwot Ridge] Now, plot the same plot but with NLCD classes separated into three facets rather than separated by color.

```
#6
#change date format
df_Niwot$collectDate<-ymd(df_Niwot$collectDate)

#plot of dry mass by date
plot_litter_color<-ggplot(subset(df_Niwot,functionalGroup=="Needles"))+
  geom_point(aes(x=collectDate,y=dryMass,color=nlcdClass))+
  ggtitle("Dry Mass of Needle Litter")+
  labs(x="Date",y="dry mass",color="NLCD class")+
  scale_x_date(limits = as.Date(c("2016-06-16","2019-09-25")),
               date_breaks = "2 months", date_labels = "%b %d %Y")+ #adjust date format
  my_plot_theme+
  theme(axis.text.x=element_text(angle=45,hjust=1,size=8)) #adjust date font size
plot_litter_color
```




```
#7
#plot of dry mass by NLDC class
plot_litter_facet<-ggplot(subset(df_Niwot,functionalGroup=="Needles"))+
  geom_point(aes(x=collectDate,y=dryMass))+
  facet_wrap(vars(nlcdClass))+
  ggtitle("Dry Mass of Needle Litter")+
  labs(x="Date",y="dry mass",color="NLCD class")+
  scale_x_date(limits = as.Date(c("2016-06-16","2019-09-25")),
               date_breaks = "6 months", date_labels = "%b %d %Y")+
  my_plot_theme+
  theme(axis.text.x=element_text(angle=45,hjust=1,size=8))
plot_litter_facet
```



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

Answer:Plot 7 is more effective in visualizing the distribution of Needle litter dry mass by NLCD class. While plot 6 differentiates the NLCD class by different colors, most data points are concentrated in the range of dry mass 0-15, and it is hard to observe the different dry mass distributions by date.