Assignment 5: Data Visualization

Shubhangi Gupta

Spring 2024

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).

2. Make sure R is reading dates as date format; if not change the format to date.

```
#1

#Verifying home directory
getwd()
```

```
## [1] "/home/guest/RStudio Project Folder/EDA_Spring2024"
```

```
#Loading packages
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
```

```
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(here)
```

```
## here() starts at /home/guest/RStudio Project Folder/EDA_Spring2024
```

```
library(cowplot)
```

```
##
## Attaching package: 'cowplot'
##
## The following object is masked from 'package:lubridate':
##
##     stamp
```

```
#Reading dataset
NTL_LTER_Lake_Chemistry_Nutrients_PeterPaul <- read.csv("Data/Processed_KEY/NTL-LTER_Lake_Chemistry_Nutr

NEON_NIWO_Litter_mass_trap <- read.csv("Data/Processed_KEY/NEON_NIWO_Litter_mass_trap_Processed.csv", st

#2
glimpse(NTL_LTER_Lake_Chemistry_Nutrients_PeterPaul)
```

```
## Rows: 23,008
## Columns: 15
## $ lakename        <fct> Paul Lake, Paul Lake, Paul Lake, Paul Lake, Paul Lake,~
## $ year4           <int> 1984, 1984, 1984, 1984, 1984, 1984, 1984, 1984, 1984, ~
## $ daynum          <int> 148, 148, 148, 148, 148, 148, 148, 148, 148, 148, 148,~
## $ month           <int> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ~
## $ sampledate      <fct> 1984-05-27, 1984-05-27, 1984-05-27, 1984-05-27, 1984-0~
## $ depth           <dbl> 0.00, 0.25, 0.50, 0.75, 1.00, 1.50, 2.00, 3.00, 4.00, ~
## $ temperature_C   <dbl> 14.5, NA, NA, NA, 14.5, NA, 14.2, 11.0, 7.0, 6.1, 5.5,~
## $ dissolvedOxygen <dbl> 9.5, NA, NA, NA, 8.8, NA, 8.6, 11.5, 11.9, 2.5, 1.6, 0~
## $ irradianceWater <dbl> 1750.0, 1550.0, 1150.0, 975.0, 870.0, 610.0, 420.0, 22~
## $ irradianceDeck  <dbl> 1620, 1620, 1620, 1620, 1620, 1620, 1620, 1620, 1620, ~
## $ tn_ug           <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ tp_ug           <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ nh34            <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ no23            <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ po4             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
```

```
NTL_LTER_Lake_Chemistry_Nutrients_PeterPaul$sampledate <- ymd(NTL_LTER_Lake_Chemistry_Nutrients_PeterPau
class(NTL_LTER_Lake_Chemistry_Nutrients_PeterPaul$sampledate)
```

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

```
glimpse(NEON_NIWO_Litter_mass_trap)
```

```
## Rows: 1,692
## Columns: 13
## $ plotID           <fct> NIWO_062, NIWO_061, NIWO_062, NIWO_064, NIWO_058, NIW~
## $ trapID           <fct> NIWO_062_050, NIWO_061_169, NIWO_062_050, NIWO_064_10~
## $ collectDate      <fct> 2016-06-16, 2016-06-16, 2016-06-16, 2016-06-16, 2016-~
## $ functionalGroup  <fct> Seeds, Other, Woody material, Seeds, Needles, Leaves,~
## $ dryMass          <dbl> 0.000, 0.270, 0.120, 0.000, 1.110, 0.000, 0.000, 0.00~
## $ qaDryMass        <fct> N, N, N, N, Y, N, N, N, N, N, N, Y, N, N, N, N, N, Y,~
## $ subplotID        <int> 31, 41, 31, 32, 32, 32, 40, 40, 40, 40, 40, 31, 31, 3~
## $ decimalLatitude  <dbl> 40.05114, 40.04762, 40.05114, 40.04737, 40.04872, 40.~
## $ decimalLongitude <dbl> -105.5858, -105.5861, -105.5858, -105.5840, -105.5872~
## $ elevation        <dbl> 3477.0, 3413.4, 3477.0, 3373.2, 3446.4, 3446.4, 3509.~
## $ nlcdClass        <fct> shrubScrub, evergreenForest, shrubScrub, evergreenFor~
## $ plotType         <fct> tower, tower, tower, tower, tower, tower, tower, towe~
## $ geodeticDatum    <fct> WGS84, WGS84, WGS84, WGS84, WGS84, WGS84, WGS84, WGS8~
```

```
NEON_NIWO_Litter_mass_trap$collectDate <- ymd(NEON_NIWO_Litter_mass_trap$collectDate)
class(NEON_NIWO_Litter_mass_trap$collectDate)
```

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

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
#Defining my theme (plot size, axes lables and legend position)
mytheme <- theme_classic(base_size = 14)+
  theme(axis.text = element_text(color="black"), legend.position = "top")

#Setting my theme as the default
theme_set(mytheme)
```

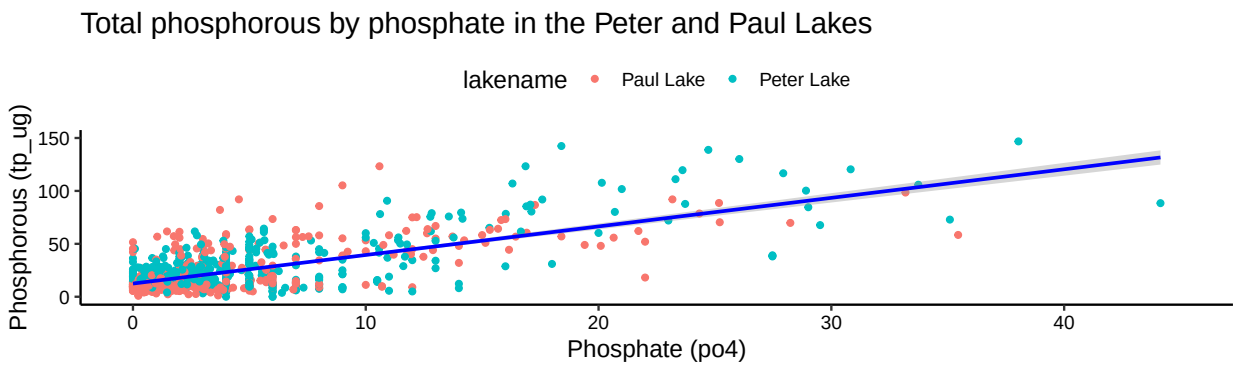
Create graphs

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 line(s) of best fit using the `lm` method. Adjust your axes to hide extreme values (hint: change the limits using `xlim()` and/or `ylim()`).

```r
#4
ggplot(NTL_LTER_Lake_Chemistry_Nutrients_PeterPaul, aes(x=po4, y=tp_ug), na.rm=TRUE )+
  geom_point(aes(color = lakename), na.rm=TRUE)+
  xlab("Phosphate (po4)")+ ylab("Phosphorous (tp_ug)")+
  ggtitle("Total phosphorous by phosphate in the Peter and Paul Lakes")+
  xlim(0,45)+ylim(0,150)+
  geom_smooth(method="lm", color = "blue", na.rm=TRUE)
```

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

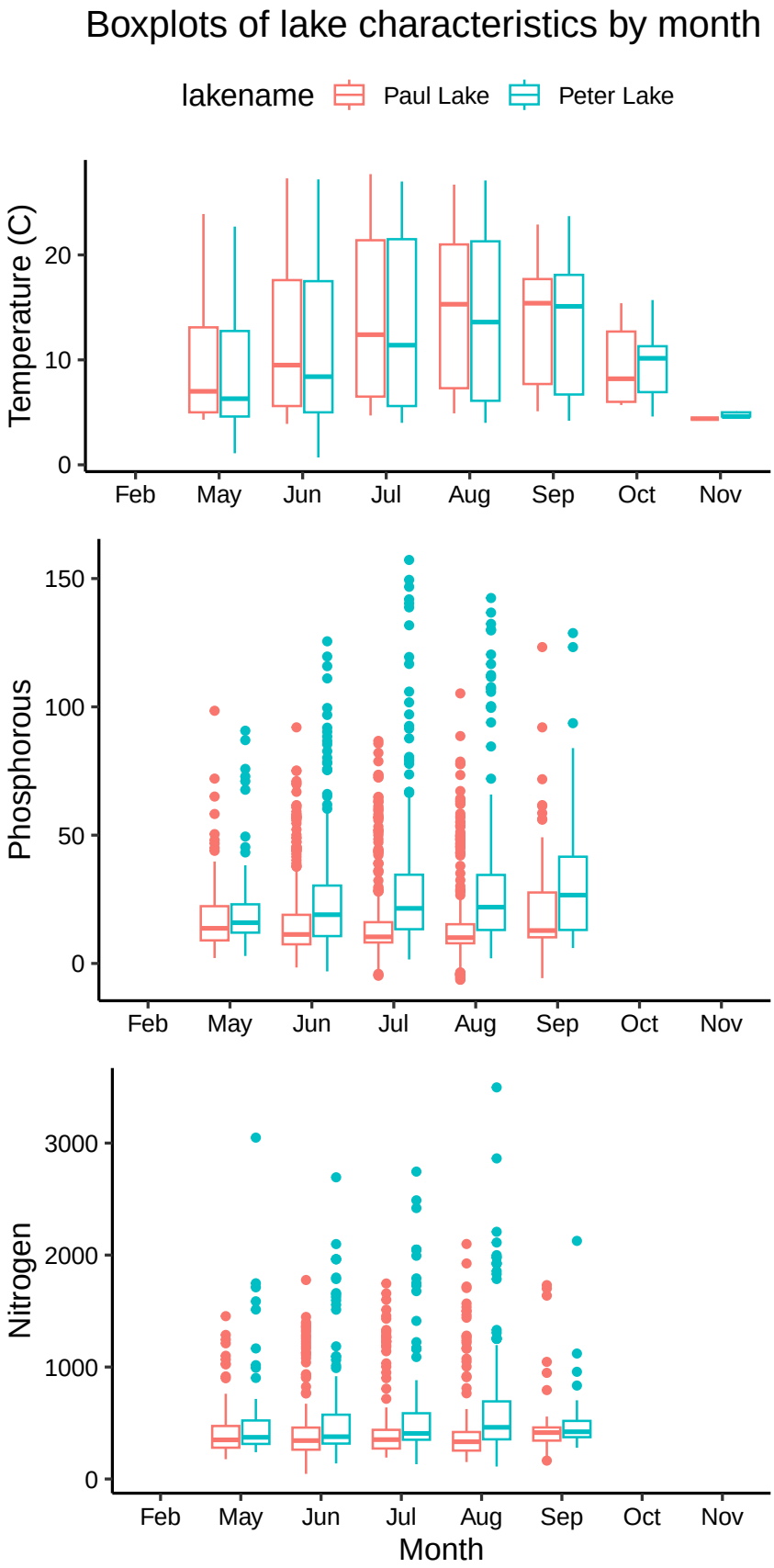


5. [NTL-LTER] Make three separate boxplots of (a) temperature, (b) TP, and (c) TN, with month as the x axis and lake as a color aesthetic. Then, create a cowplot that combines the three graphs. Make sure that only one legend is present and that graph axes are aligned.

Tips: * Recall the discussion on factors in the lab section as it may be helpful here. * Setting an axis title in your theme to `element_blank()` removes the axis title (useful when multiple, aligned plots use the same axis values) * Setting a legend's position to "none" will remove the legend from a plot. * Individual plots can have different sizes when combined using `cowplot`.

```r
#5

#Adding a new column to the df with the months as factors
NTL_LTER_Lake_Chemistry_Nutrients_PeterPaul$month_factor <- factor(NTL_LTER_Lake_Chemistry_Nutrients_Pe

#Plotting the three boxplots separately
temp_plot <- ggplot(NTL_LTER_Lake_Chemistry_Nutrients_PeterPaul, aes(x=month_factor, y=temperature_C), n
  geom_boxplot(aes(color = lakename), na.rm=TRUE)+
 ylab("Temperature (C)")+
  theme(axis.title.x = element_blank())+
  ggtitle("Boxplots of lake characteristics by month")


tp_plot <- ggplot(NTL_LTER_Lake_Chemistry_Nutrients_PeterPaul, aes(x=month_factor, y=tp_ug), na.rm=TRUE
  geom_boxplot(aes(color = lakename), na.rm=TRUE)+
  ylab("Phosphorous")+
  theme(legend.position = "none", axis.title.x = element_blank())

TN_plot <- ggplot(NTL_LTER_Lake_Chemistry_Nutrients_PeterPaul, aes(x=month_factor, y=tn_ug), na.rm=TRUE
  geom_boxplot(aes(color = lakename), na.rm=TRUE)+
```

```
  ylab("Nitrogen")+ xlab("Month")+
  theme(legend.position = "none")

#Combining the three plots into one grid
plot_grid(temp_plot, tp_plot, TN_plot, nrow=3, rel_widths = c(2,2,2))
```


Boxplots of lake characteristics by month
lakename
Paul Lake
Peter Lake
Temperature (C)
0
10
20
Feb
May
Jun
Jul
Aug
Sep
Oct
Nov
Phosphorous
0
50
100
150
Nitrogen
0
1000
2000
3000
Month

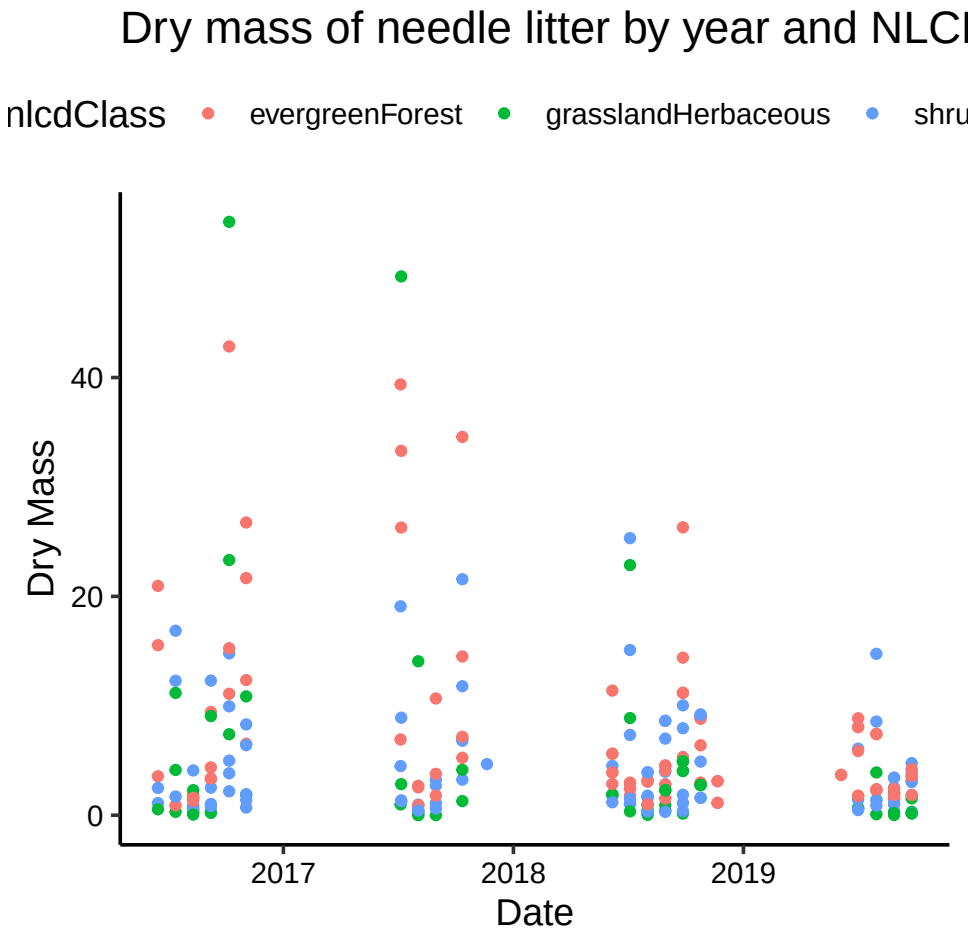
Question: What do you observe about the variables of interest over seasons and between lakes?

> Answer: The temperature of both lakes rises from May to September and then declines in October. Paul lake is warmer than Peter Lake during all the summer months, and cooler than it during October. On the contrary, phosphorous levels in Paul lake decline from May to August and then rise slightly in September, while that of Peter lake rise from May to September. Both lakes have a considerable number of outliers in this phosphorous data that reach very high values in the summer months compared to the median. Nitrogen levels remain fairly constant across timein both lakes, although that of Peter lake is slightly higher than that of Paul lake. The nitrogen data also consists of a fair amount of outliers compared to the median, especially in Paul Lake.

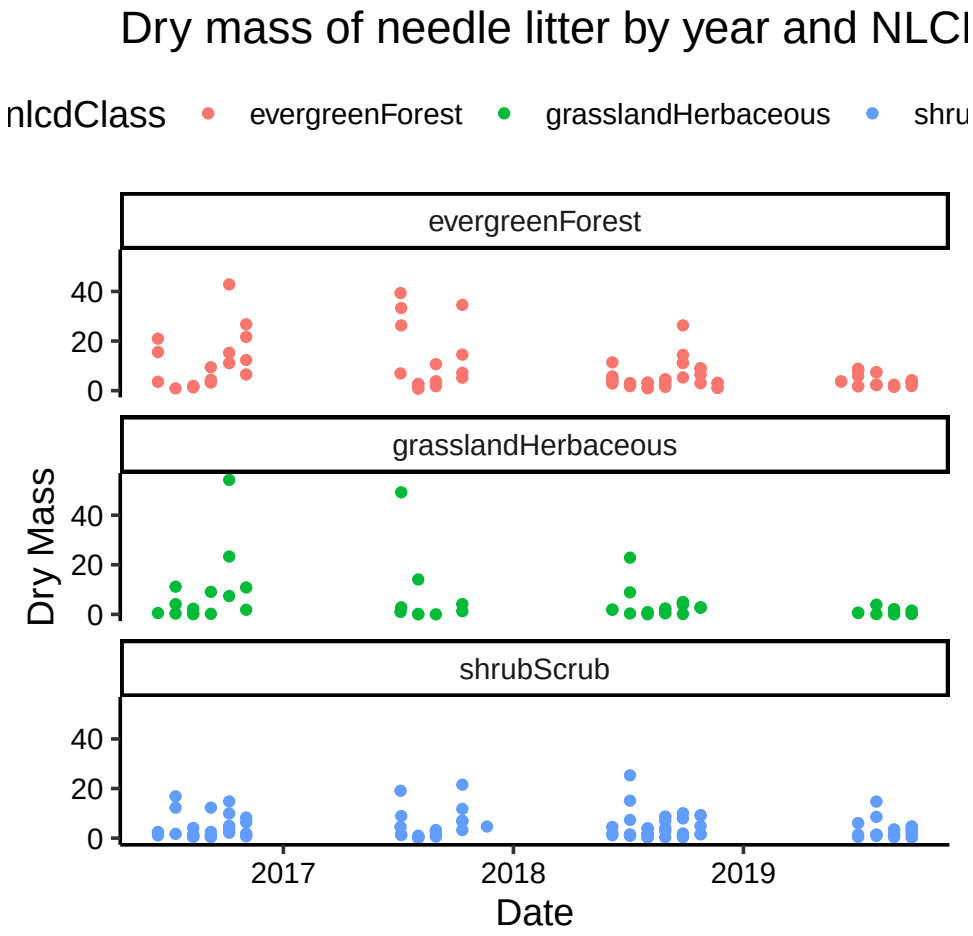
6. [Niwot Ridge] Plot a subset of the litter dataset by displaying only the "Needles" functional group. Plot the dry mass of needle litter by date and separate by NLCD class with a color aesthetic. (no need to adjust the name of each land use)

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

```
#6

ggplot(subset(NEON_NIWO_Litter_mass_trap, functionalGroup=="Needles"), na.rm=TRUE)+
  geom_point(aes(x=collectDate, y=dryMass, color = nlcdClass), na.rm=TRUE)+
  xlab("Date")+ylab("Dry Mass")+
  ggtitle("Dry mass of needle litter by year and NLCD class")
```




```
#7
ggplot(subset(NEON_NIWO_Litter_mass_trap, functionalGroup=="Needles"), na.rm=TRUE)+
  geom_point(aes(x=collectDate, y=dryMass, color=nlcdClass), na.rm=TRUE)+
  facet_wrap(vars(nlcdClass), nrow = 3)+
  xlab("Date")+ylab("Dry Mass")+
  ggtitle("Dry mass of needle litter by year and NLCD class")
```



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

Answer:7 is more effective because it separates out the data points for the three nlcd classes making it possible for us to observe their trends and distribution more distinctly.