Assignment 5: Data Visualization

Aishwarya Patankar

Fall 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 Loading the required packages and getting the working directory
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## 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
```

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

```
## here() starts at /Users/aishwaryapatankar/Documents/Duke University/Spring2025/ENERGY872/EDA_Spring20
```

```r
#install.packages("cowplot") : Hashed out while knittimg file
library(cowplot)
```

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

```r
getwd()
```

```
## [1] "/Users/aishwaryapatankar/Documents/Duke University/Spring2025/ENERGY872/EDA_Spring2025"
```

```r
ChemistryNutrients <- read.csv(
  file = here("./Data/Processed_KEY/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv"),
  stringsAsFactors = TRUE)

NiwotRidge_Litter <- read.csv(
  file = here("./Data/Processed_KEY/NEON_NIWO_Litter_mass_trap_Processed.csv"),
  stringsAsFactors = TRUE)

#2
class(ChemistryNutrients$sampledate)
```

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

```r
class(NiwotRidge_Litter$collectDate)
```

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

```r
ChemistryNutrients$sampledate<- as.Date(ChemistryNutrients$sampledate, format = "%Y-%m-%d")
NiwotRidge_Litter$collectDate<- as.Date(NiwotRidge_Litter$collectDate, format = "%Y-%m-%d")

class(ChemistryNutrients$sampledate)
```

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

```r
class(NiwotRidge_Litter$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
mytheme<- theme_grey(base_size = 14)+
  theme(legend.position = "top") #Setting theme with background and legend position
```

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()`).

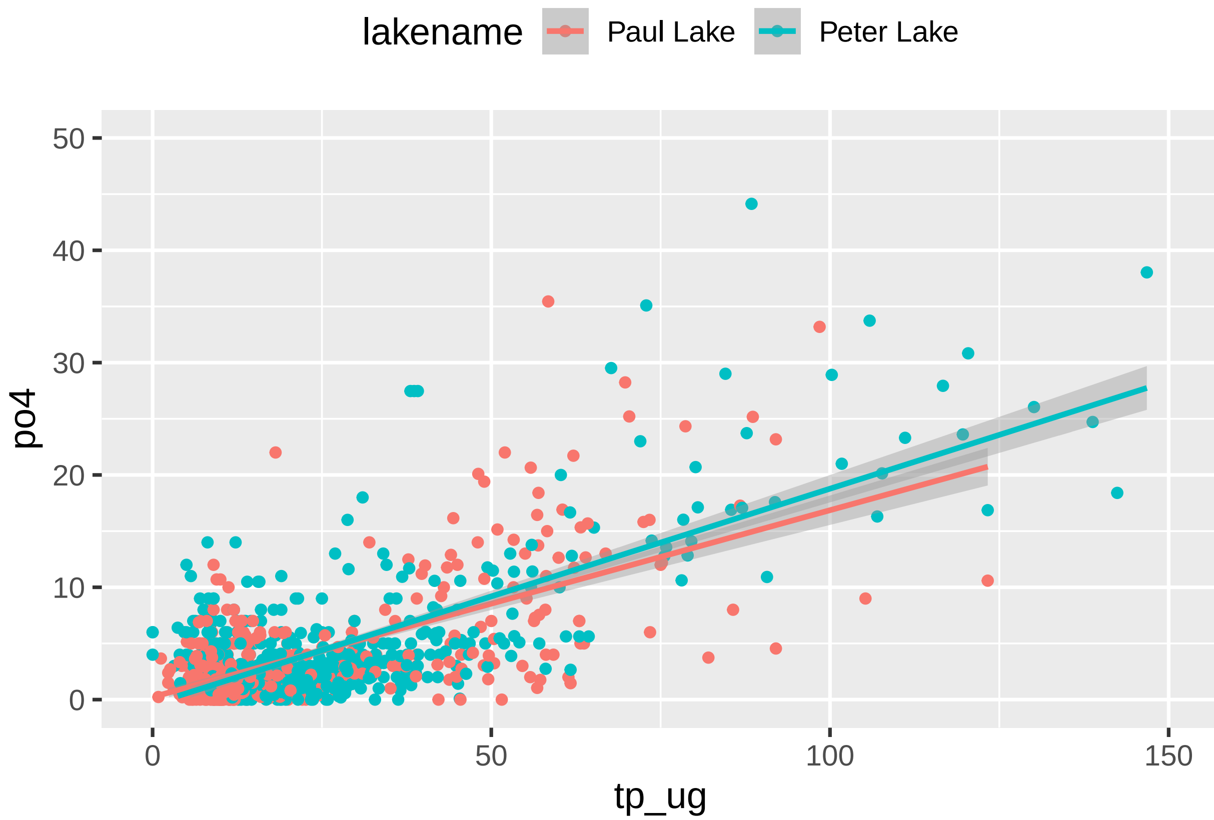
```
#4
Plot1<-
  ggplot(ChemistryNutrients, aes(x= tp_ug ,y= po4, color = lakename))+
  geom_point()+
  geom_smooth(method = lm)+
  xlim(0,150)+
  ylim(0,50)+
  mytheme
  print(Plot1)
```

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

## Warning: Removed 21948 rows containing non-finite outside the scale range
## (`stat_smooth()`).

## Warning: Removed 21948 rows containing missing values or values outside the scale range
## (`geom_point()`).

## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`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.

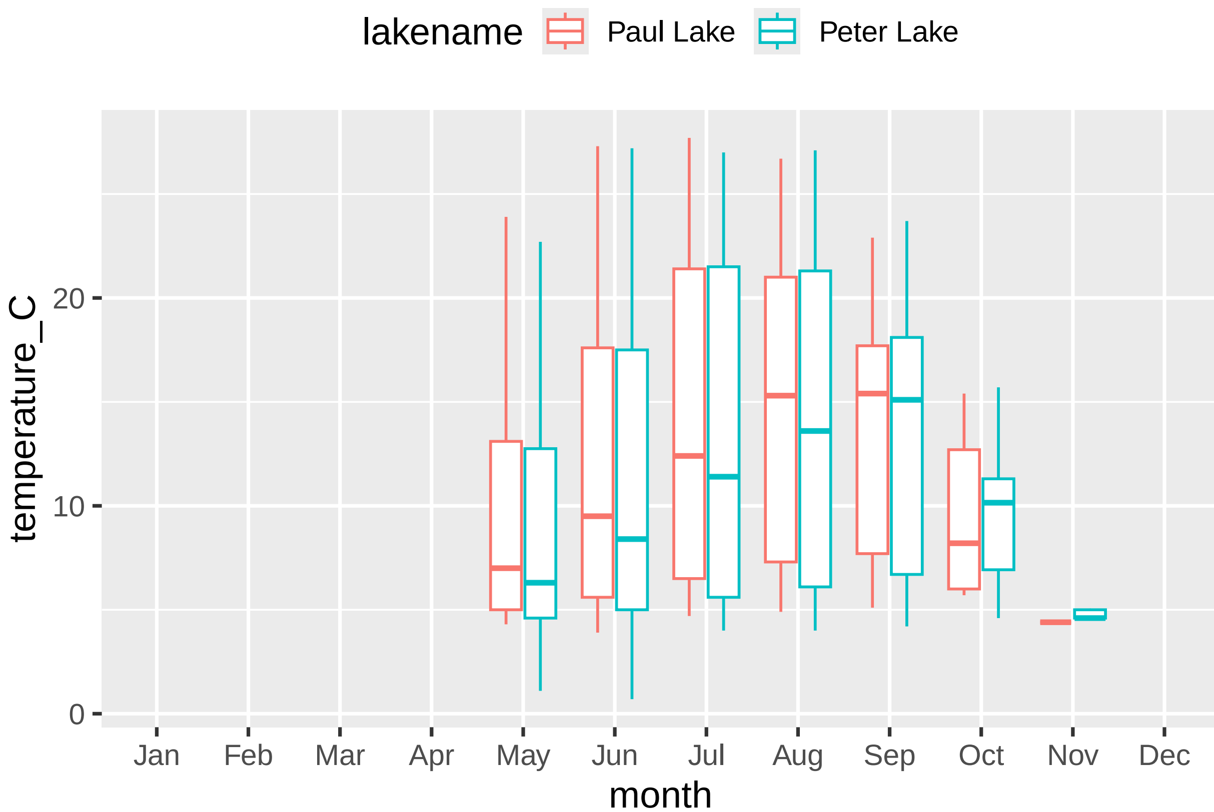
Tips: * Recall the discussion on factors in the lab section as it may be helpful here. * Setting an axis title in your theme to `element_blank()` removes the axis title (useful when multiple, aligned plots use the same axis values) * Setting a legend's position to "none" will remove the legend from a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

```
#5
class(ChemistryNutrients$month)
```

```
## [1] "integer"
```

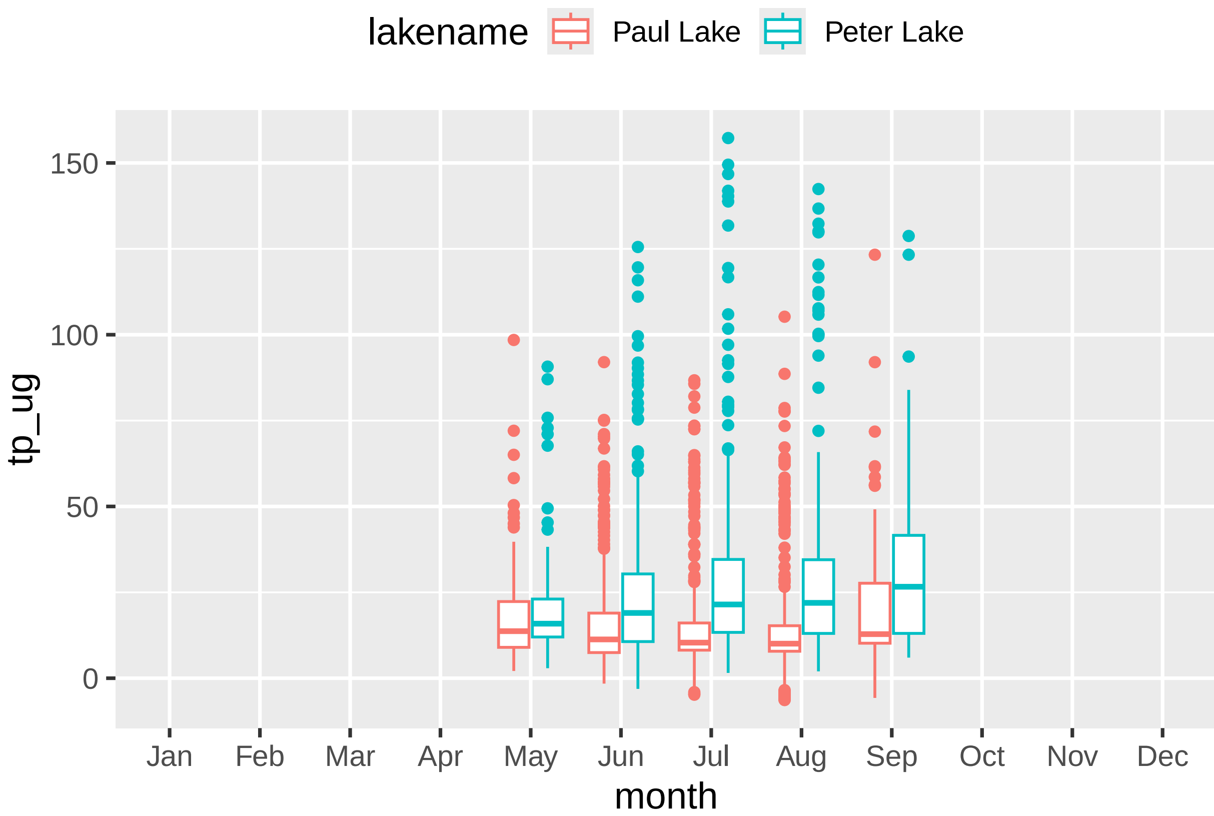
```
Plot2a<-
  ggplot(ChemistryNutrients, aes(x =factor(month,levels = 1:12, labels=month.abb )))+
  geom_boxplot(aes(y = temperature_C, color = lakename))+
  scale_x_discrete(name = 'month', drop=FALSE)+
  mytheme
  print(Plot2a)
```

```
## Warning: Removed 3566 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```



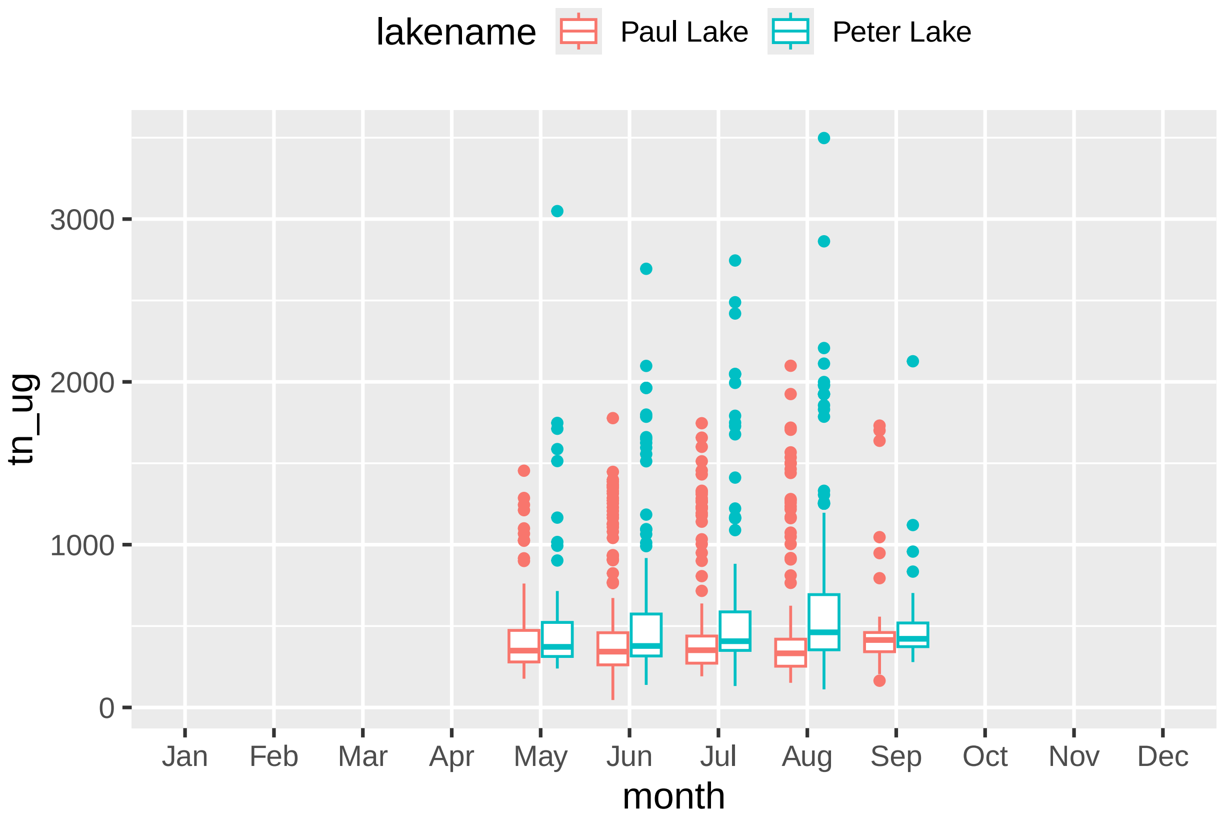

```
Plot2b<-
  ggplot(ChemistryNutrients, aes(x =factor(month,levels = 1:12, labels=month.abb )))+
  geom_boxplot(aes(y = tp_ug, color = lakename))+
  scale_x_discrete(name = 'month', drop=FALSE)+
  mytheme
  print(Plot2b)
```

```
## Warning: Removed 20729 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```




```
Plot2c<-
  ggplot(ChemistryNutrients, aes(x =factor(month,levels = 1:12, labels=month.abb )))+
  geom_boxplot(aes(y = tn_ug, color = lakename))+
  scale_x_discrete(name = 'month', drop=FALSE)+
  mytheme
  print(Plot2c)
```

```
## Warning: Removed 21583 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



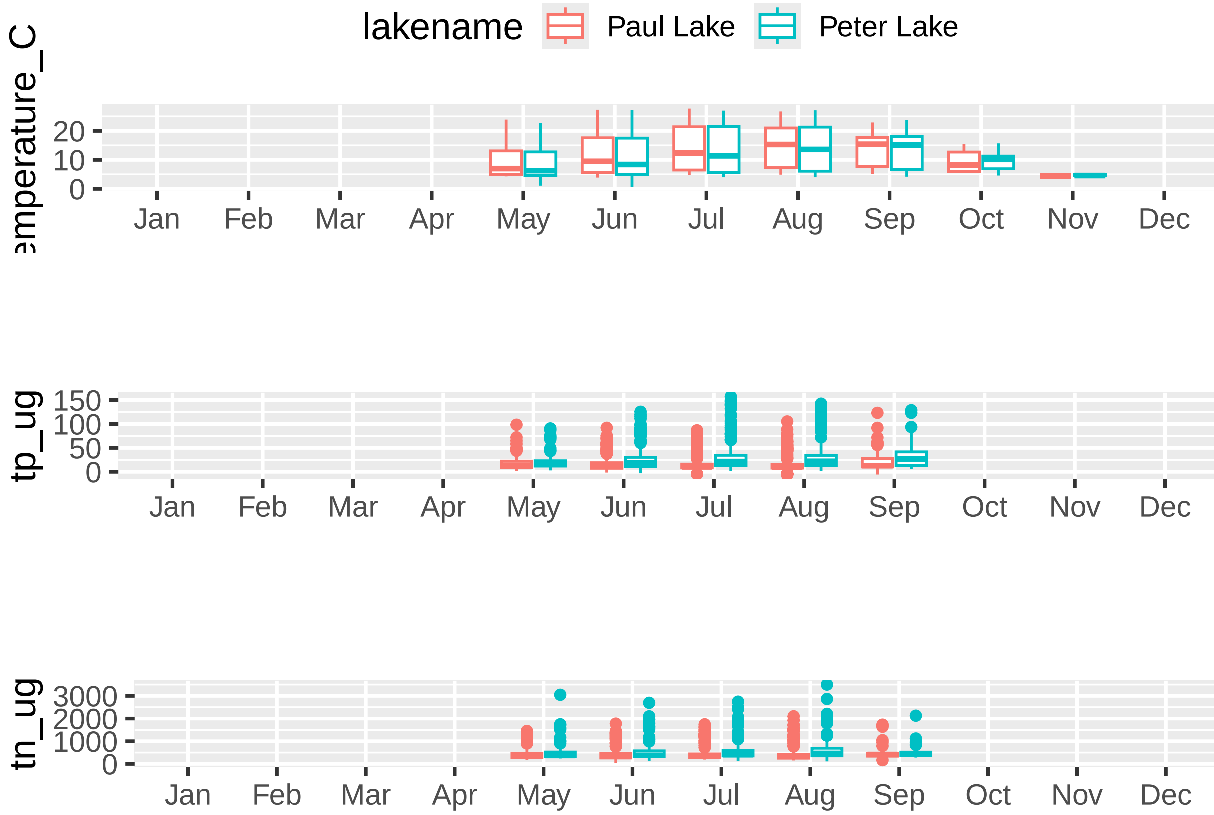

```r
#installing cowplot
#install.packages("cowplot") : Line is hashed out while knitting
library(cowplot)
#Creating a combine plot and removing the legends from plot2b and 2c and removing axis titles from all
combinedplot <- plot_grid(Plot2a + theme(axis.title.x = element_blank()),Plot2b + theme(axis.title.x = e
```

```
## Warning: Removed 3566 rows containing non-finite outside the scale range
## (`stat_boxplot()`).

## Warning: Removed 20729 rows containing non-finite outside the scale range
## (`stat_boxplot()`).

## Warning: Removed 21583 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```

```r
print(combinedplot)
```



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

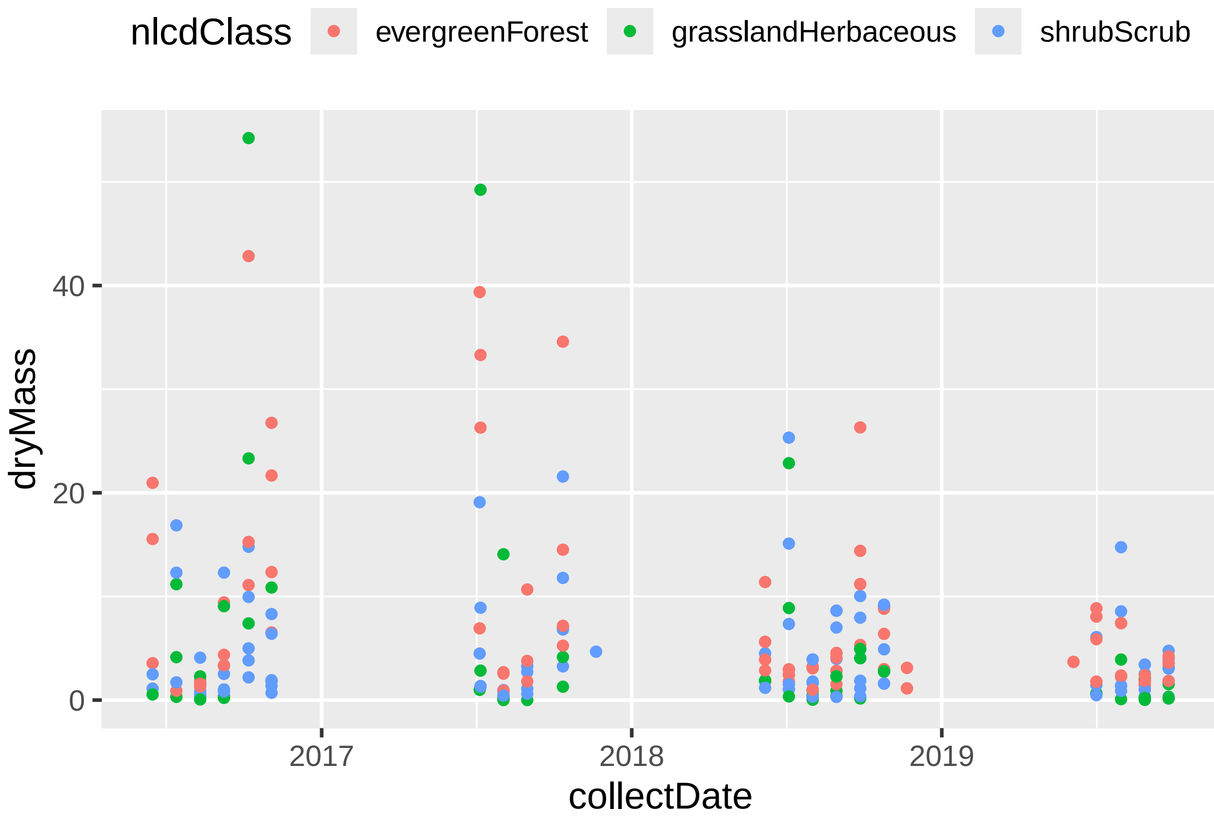
> Answer: We observe that the temperature fluctuations are higher in months of June, July and August and taper off towards winter as seen in the November plot. Both Paul Lake and Peter Lake have similar temperature profile. The Total Phosphorus content is higher in Peter Lake as compared to Paul Lake and difference from mean of the phosphorus content is also higher in Peter Lake as compared to Paul Lake. Similarly, the Total Nitrogen content is higher in Peter Lake as compared to Paul Lake and difference from mean of the nitrogen content is also higher in Peter Lake as compared to 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.

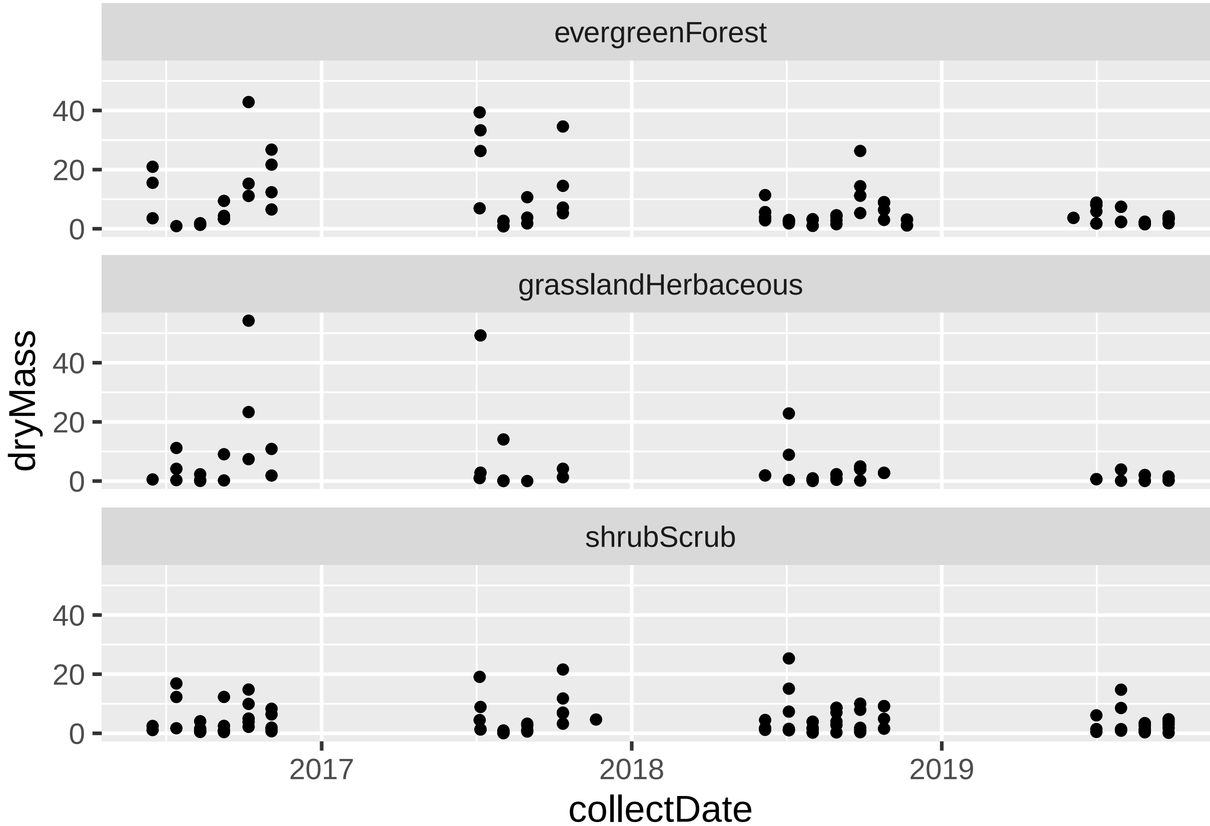
```
#6
#Filtering the Functional Group 'Needles'
Niwot_Needle<-
  NiwotRidge_Litter%>%
  filter(functionalGroup == "Needles")

Plot3<-
  ggplot(Niwot_Needle, aes(x = collectDate, y = dryMass, color = nlcdClass))+
  geom_point()+
  mytheme
  print(Plot3)
```




```r
#7

Plot4<-
  ggplot(Niwot_Needle, aes(x = collectDate, y = dryMass))+
  geom_point()+
  facet_wrap(vars(nlcdClass), nrow = 3)+
  mytheme
  print(Plot4)
```



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

> Answer: I think the plot prepared for Q6 is more effective because it allows us to compare the drymass of different NLCD Classes with each other in a single frame and make inferences about Increase or decrease in dry mass for different classes and across different years.