Assignment 5: Data Visualization

Jackie Fahrenholz

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
#verify working directory
print(getwd())
```

```
## [1] "C:/Users/Jackie/Desktop/ENV872/Environmental_Data_Analytics_2022/Assignments"
```

```
#load packages
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)
#upload datasets #these run off the page because it wouldn't knit otherwise
chem.nutrients <- read.csv("../Data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv
litter <- read.csv("../Data/Processed/NEON_NIWO_Litter_mass_trap_Processed.csv", stringsAsFactors = TRU
```

```
#2
#check dates for format
class(chem.nutrients$sampledate)
```

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

```
#not in date so change it
chem.nutrients$sampledate <- as.Date(chem.nutrients$sampledate,
                                     format = "%Y-%m-%d")
#check it again
class(chem.nutrients$sampledate)
```

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

```
#check other dataset for date format
class(litter$collectDate)
```

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

```
#change this one too
litter$collectDate <- as.Date(litter$collectDate, format = "%Y-%m-%d")
#check it
class(litter$collectDate)
```

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

Define your theme

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

```
#3
#lets build a theme!
mytheme <- theme_classic(base_size = 13) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "right")
```

Create graphs

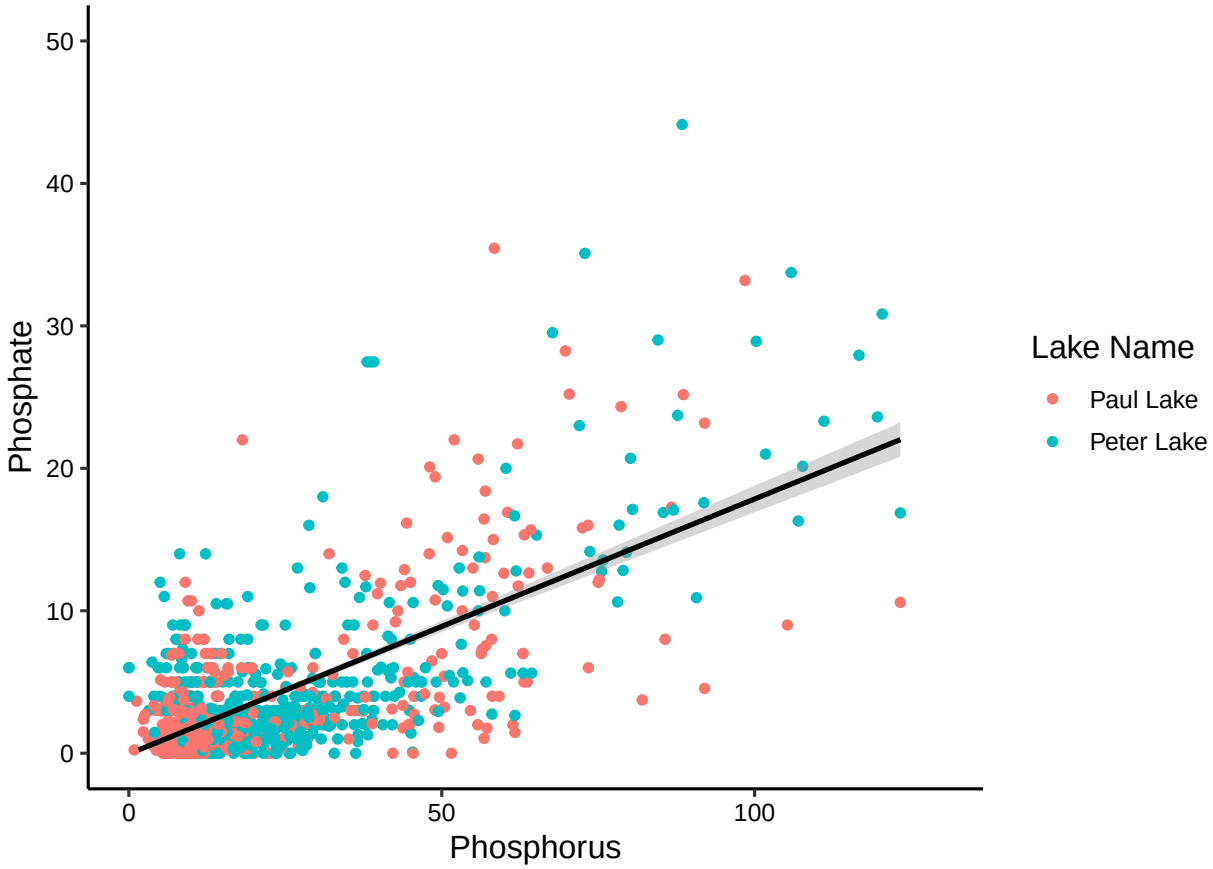
For numbers 4-7, create ggplot graphs and adjust aesthetics to follow best practices for data visualization. Ensure your theme, color palettes, axes, and additional aesthetics are edited accordingly.

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 `ylim()`).

```
#4
#set theme
theme_set(mytheme)
#plot this
#define x, y and color by lakename
ggplot(chem.nutrients, aes(x = tp_ug, y = po4, color = lakename))+
  geom_point()+
  #line of best fit, colored black
  geom_smooth(method = "lm", color = "black")+
  #choose your limits; removing anymore on the x-axis wouldn't be
  #representative, despite there being a clump of data between 0-50
  ylim(0,50) +
  xlim(0,130) +
```

```
  ylab('Phosphate') +
  xlab('Phosphorus')+
  labs(color = "Lake Name")
```

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

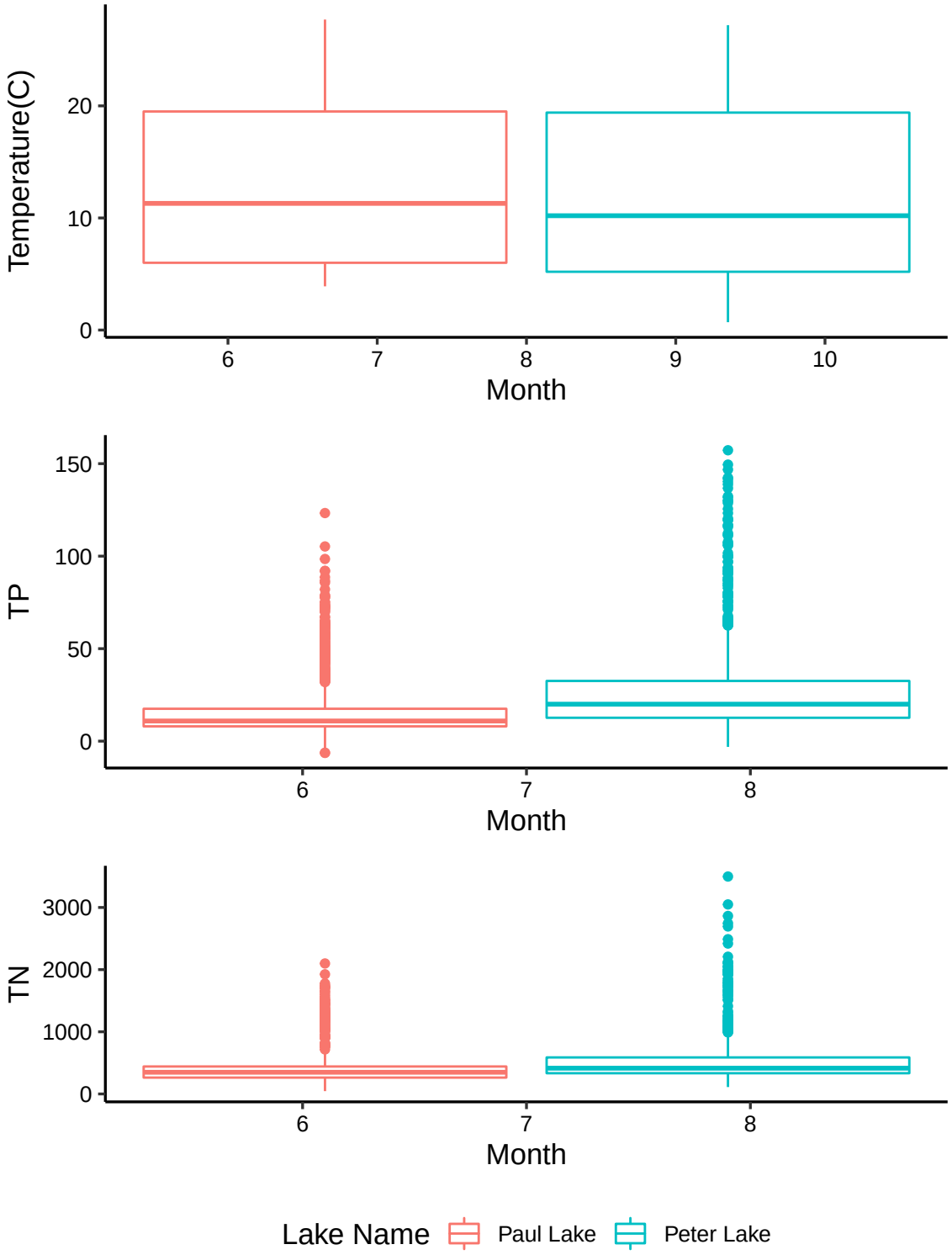


5. [NTL-LTER] Make three separate boxplots of (a) temperature, (b) TP, and (c) TN, with month as the x axis and lake as a color aesthetic. Then, create a cowplot that combines the three graphs. Make sure that only one legend is present and that graph axes are aligned.

```
#5
#set theme
theme_set(mytheme)
#make temp boxplot
temp.plot <- ggplot(chem.nutrients, aes(x = month, y = temperature_C))+
  geom_boxplot(aes(color = lakename))+
  xlab('Month')+
  ylab('Temperature(C)')+
  labs(color = "Lake Name")
# TP plot
tp.plot <- ggplot(chem.nutrients, aes(x = month, y = tp_ug))+
  geom_boxplot(aes(color = lakename))+
  xlab('Month')+
  ylab('TP')+
  labs(color = "Lake Name")
# TN plot
```

```
tn.plot <- ggplot(chem.nutrients, aes(x = month, y = tn_ug))+
  geom_boxplot(aes(color = lakename))+
  xlab('Month')+
  ylab('TN')+
  labs(color = "Lake Name")

#use cowplot to make them into one figure
plot_grid(temp.plot + theme(legend.position="none"),
          tp.plot + theme(legend.position="none"),
          tn.plot  + theme(legend.position="bottom"),
          nrow = 3, align = 'hv', rel_heights = c(1, 1))
```



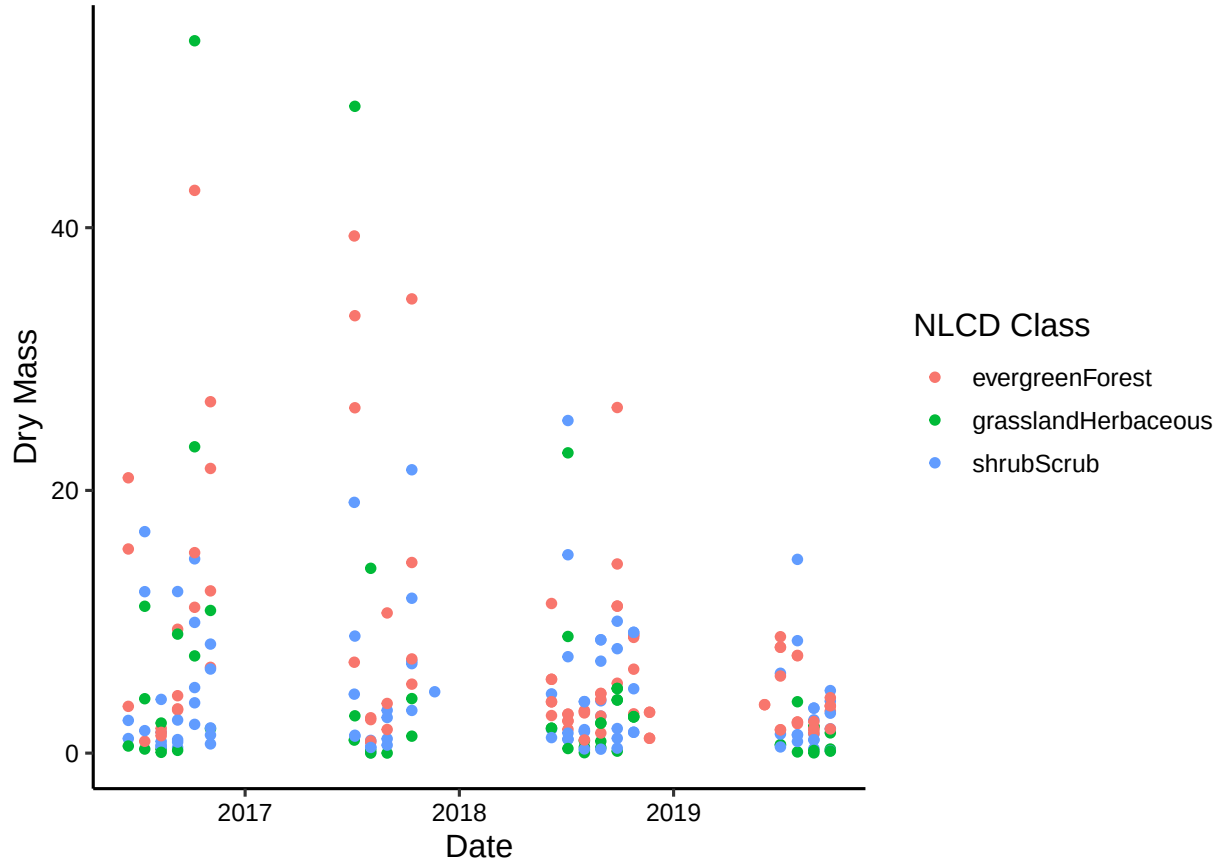
Question: What do you observe about the variables of interest over seasons and between lakes?

Answer: TP and TN were only measured during three months (6,7,8) while the temeperature was measured during all months. This may lead us to believe that you can only sample the other variables during summer months, or that this pollutant is present during this time based on seasonal activities. Temperature on average is higher for Paul lake is higher than Peter 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)

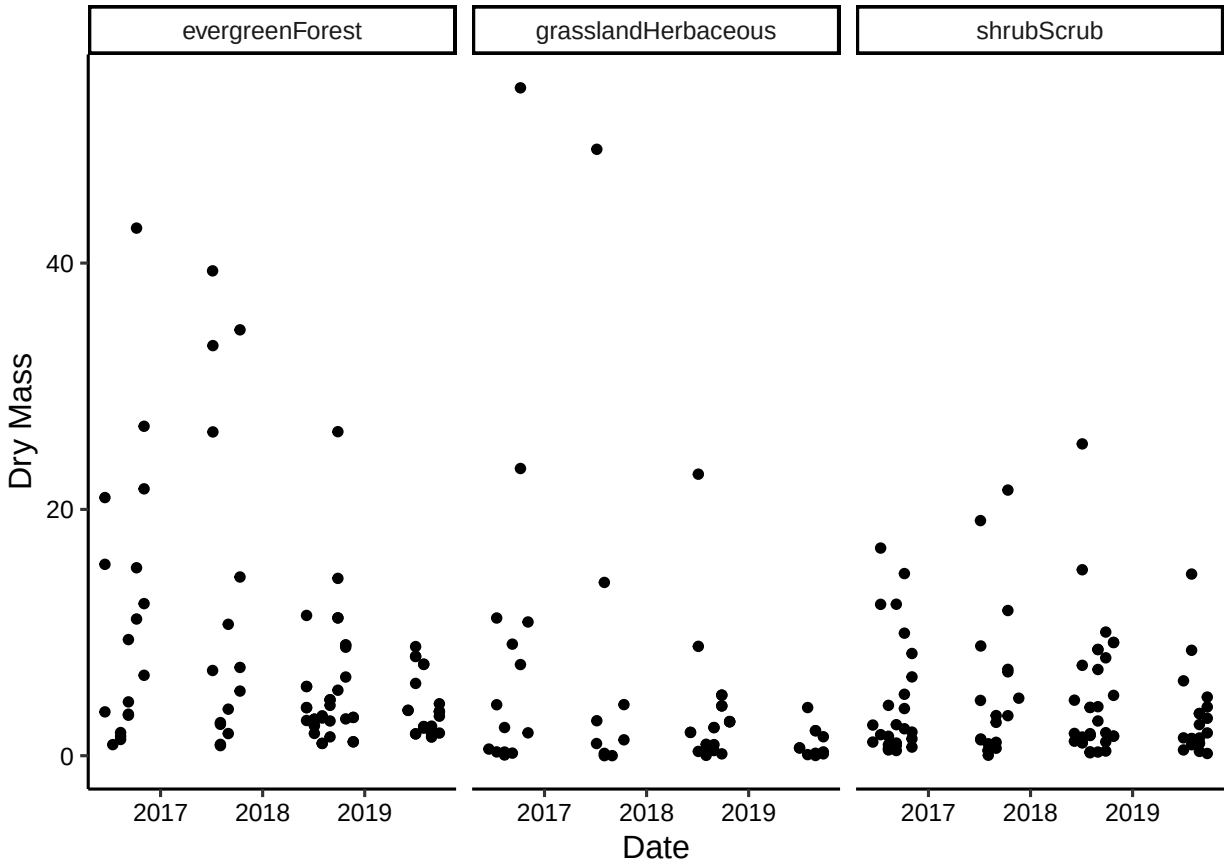
7. [Niwot Ridge] Now, plot the same plot but with NLCD classes separated into three facets rather than separated by color.

```
#6
#set theme
theme_set(mytheme)
#plot this by subsetting the litter dataset into Needles only
litter.subset <- ggplot(subset(litter, functionalGroup == "Needles"),
                        aes(x = collectDate, y = dryMass,color = nlcdClass)) +
  geom_point() +
  xlab("Date") +
  ylab("Dry Mass") +
  labs(color = "NLCD Class")
print(litter.subset)
```

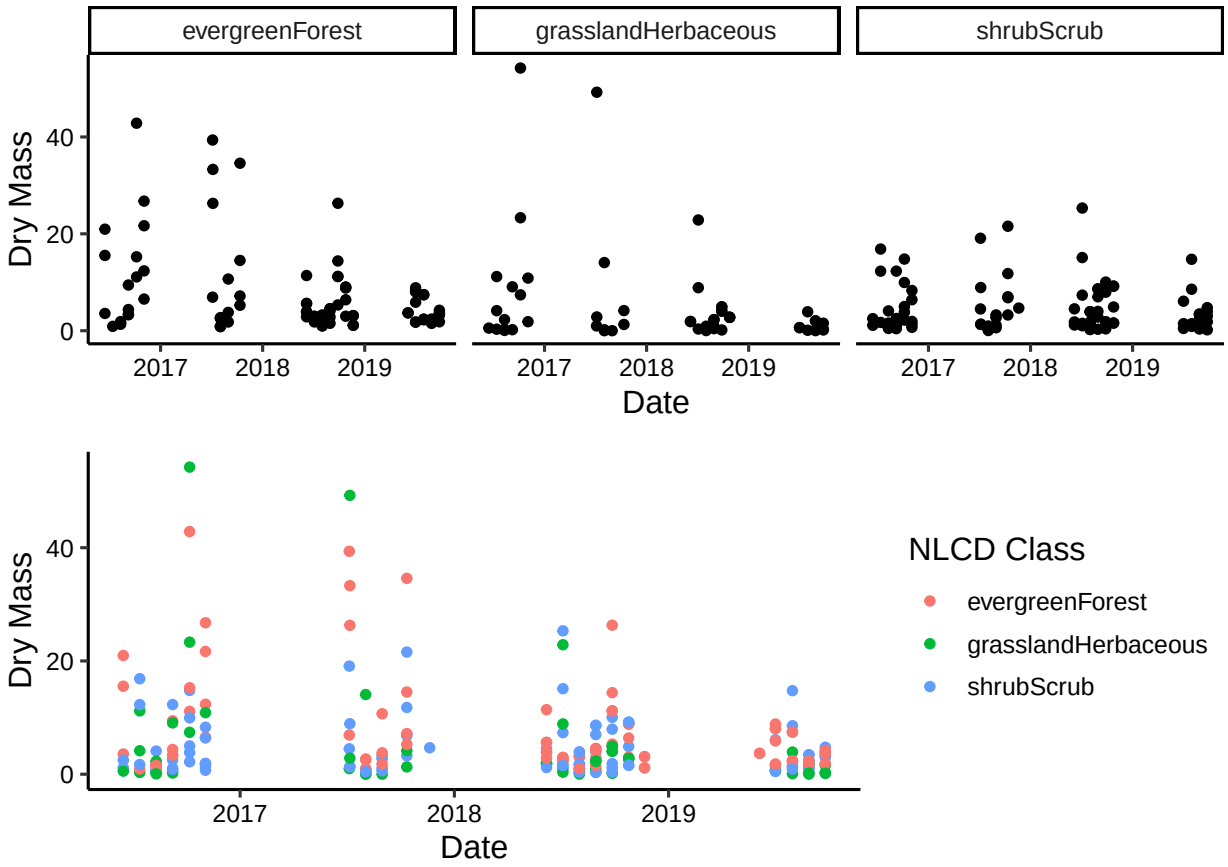



```
#7
#now do the same thing but with a facet instead of by color
litter.facet <- ggplot(subset(litter, functionalGroup == "Needles"),
                       aes(x = collectDate, y = dryMass)) +
  facet_wrap(vars(nlcdClass), ncol = 3) +
  geom_point() +
  xlab("Date") +
  ylab("Dry Mass")
```

```
print(litter.facet)
```




```
# lets put them next to each other to see what we think
plot_grid(litter.facet, litter.subset, nrow = 2, align = 'hv',
          rel_heights = c(1, 1))
```



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

Answer: I think that the facetwrap graph `litter.facet` is more effective in displaying the differences between each of the classes because you can visually compare the density of points next to each other as well as the variation between years and types of nlcd class across the board. The first graph though it's visually appealing doesn't give us as full of a picture but rather displays how similar or different they are based on ncld class in the given year. These plots are more crowded though so only general conclusions could be made based on this scale.