Assignment 5: Data Visualization

Sena McCrory

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., "Salk_A05_DataVisualization.Rmd") prior to submission.

The completed exercise is due on Tuesday, February 11 at 1:00 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 (tidy and gathered) and the processed data file for the Niwot Ridge litter dataset.
2. Make sure R is reading dates as date format; if not change the format to date.

```
#1
getwd()
```

```
## [1] "C:/Users/senam/Box Sync/My Documents/MEM classes/Duke Spring 2020/DataAnalytics/Environmental_Da
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(cowplot)
```

```
##
## ********************************************************

## Note: As of version 1.0.0, cowplot does not change the

##   default ggplot2 theme anymore. To recover the previous
```

```
##   behavior, execute:
##   theme_set(theme_cowplot())
## ********************************************************
```

```r
library(lubridate)
```

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

```r
library(viridis)
```

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

```r
PP.nutrients.spread <- read.csv("./Data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed
PP.nutrients.gathered <- read.csv("./Data/Processed/NTL-LTER_Lake_Nutrients_PeterPaulGathered_Processed
Litter <- read.csv("./Data/Processed/NEON_NIWO_Litter_mass_trap_Processed.csv")


#2
class(PP.nutrients.spread$sampledate)
```

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

```r
class(Litter$collectDate)
```

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

```r
PP.nutrients.spread$sampledate <- as.Date(
  PP.nutrients.spread$sampledate, format = "%Y-%m-%d")
PP.nutrients.gathered$sampledate <- as.Date(
  PP.nutrients.gathered$sampledate, format = "%Y-%m-%d")
Litter$collectDate <- as.Date(
  Litter$collectDate, format = "%Y-%m-%d")
class(PP.nutrients.spread$sampledate)
```

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

```r
class(Litter$collectDate)
```

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

Define your theme

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

```r
mytheme <- theme_minimal(base_size = 12) +
  theme(axis.text = element_text(color = "gray2"),
        legend.position = "bottom")
theme_set(mytheme)
```

Create graphs

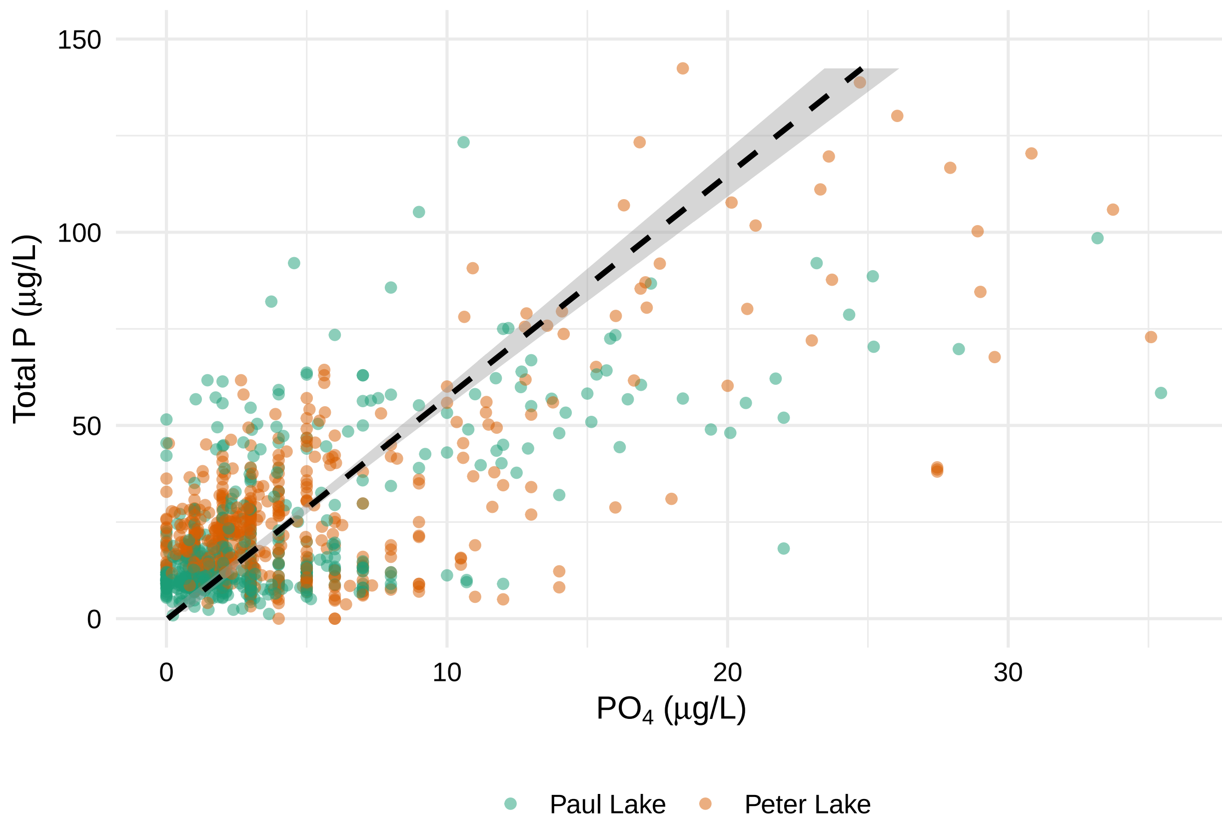
For numbers 4-7, create ggplot graphs and adjust aesthetics to follow best practices for data visualization. Ensure your theme, color palettes, axes, and additional aesthetics are edited accordingly.

4. [NTL-LTER] Plot total phosphorus by phosphate, 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.

```
ntlplot1 <- ggplot(PP.nutrients.spread)+
  geom_point(aes(x = tp_ug, y = po4, color = lakename), alpha = 0.5, size = 1.5)+
  geom_smooth(aes(x = tp_ug, y= po4), method = "lm", color = "black", linetype = 2)+
  ylim(c(0, 36))+ # three outliers not shown
  xlim(c(0, 150))+
  #scale_y_log10()+
  #scale_x_log10()+
  labs(color = "")+
  xlab(expression(paste("Total P (", mu, "g/L)")))+
  ylab(expression(paste("PO"[4]*" (",mu,"g/L)")))+
  scale_color_brewer(type = "qual", palette = 2)+
  coord_flip()
print(ntlplot1)
```

```
## Warning: Removed 21950 rows containing non-finite values (stat_smooth).
```

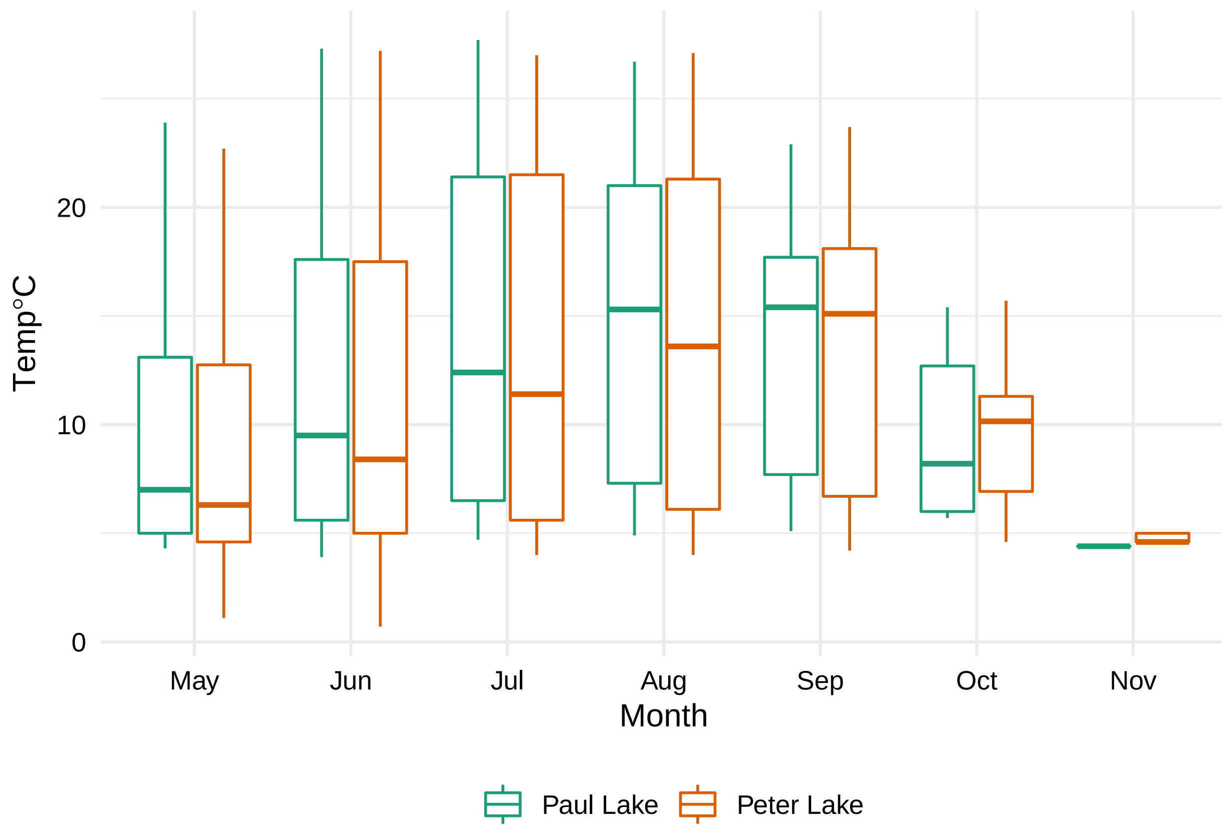
```
## Warning: Removed 21950 rows containing missing values (geom_point).
```



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.

```r
#PP.nutrients.spread$month <- as.Date(PP.nutrients.spread$month, format = "%m", origin="1970-01-01")

PP.nutrients.spread$month <- format(PP.nutrients.spread$sampledate, format = "%b")

boxplotA <- ggplot(PP.nutrients.spread)+
  geom_boxplot(aes(x = month, y = temperature_C, color = lakename))+
  labs(x = "Month", y = expression(Temp *degree* C),
       color = "")+
  scale_x_discrete(limits = month.abb[5:11])+
  scale_color_brewer(type = "qual", palette = 2)

print(boxplotA)
```

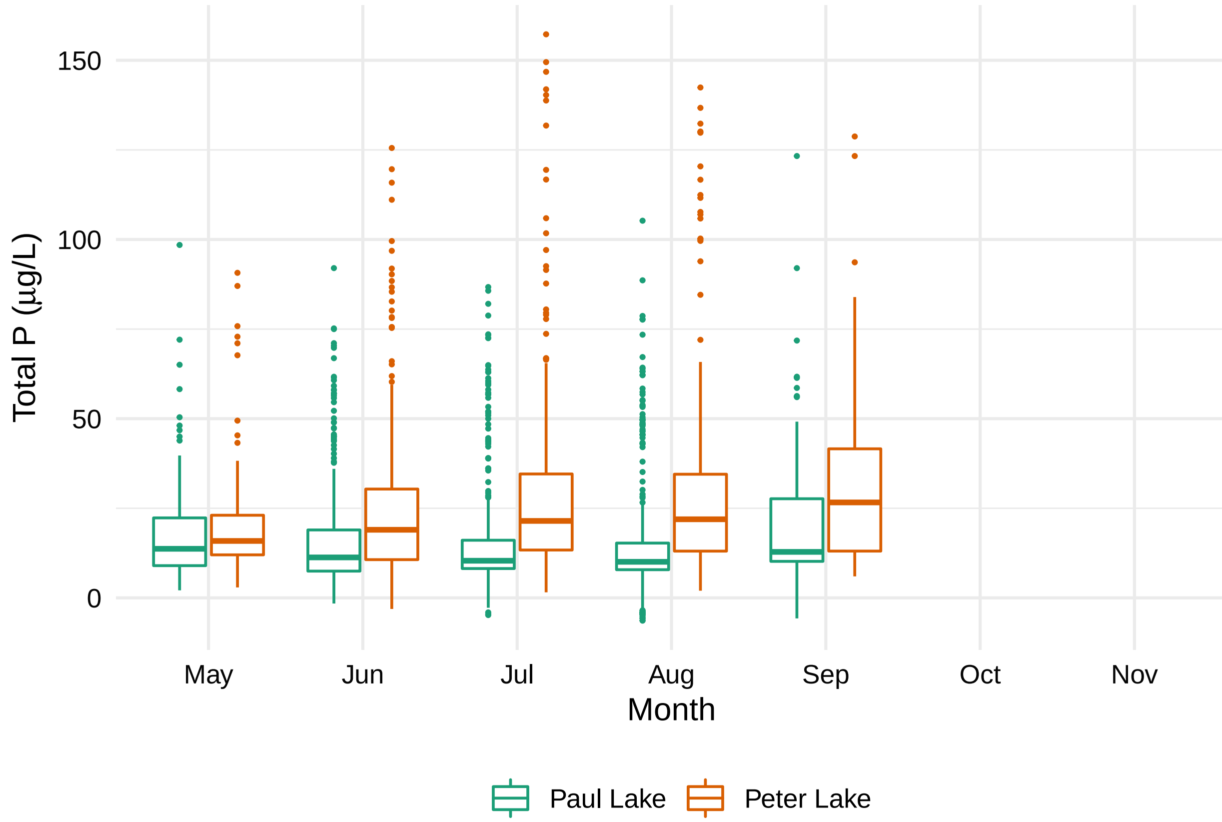
```
## Warning: Removed 16 rows containing missing values (stat_boxplot).
## Warning: Removed 3550 rows containing non-finite values (stat_boxplot).
```




```r
boxplotB <- ggplot(PP.nutrients.spread)+
  geom_boxplot(aes(x = month, y = tp_ug, color = lakename), outlier.size = 0.4)+
  labs(x = "Month", y = expression(paste("Total P (", mu, "g/L)")),
       color = "")+
  scale_x_discrete(limits = month.abb[5:11])+
  scale_color_brewer(type = "qual", palette = 2)
print(boxplotB)
```

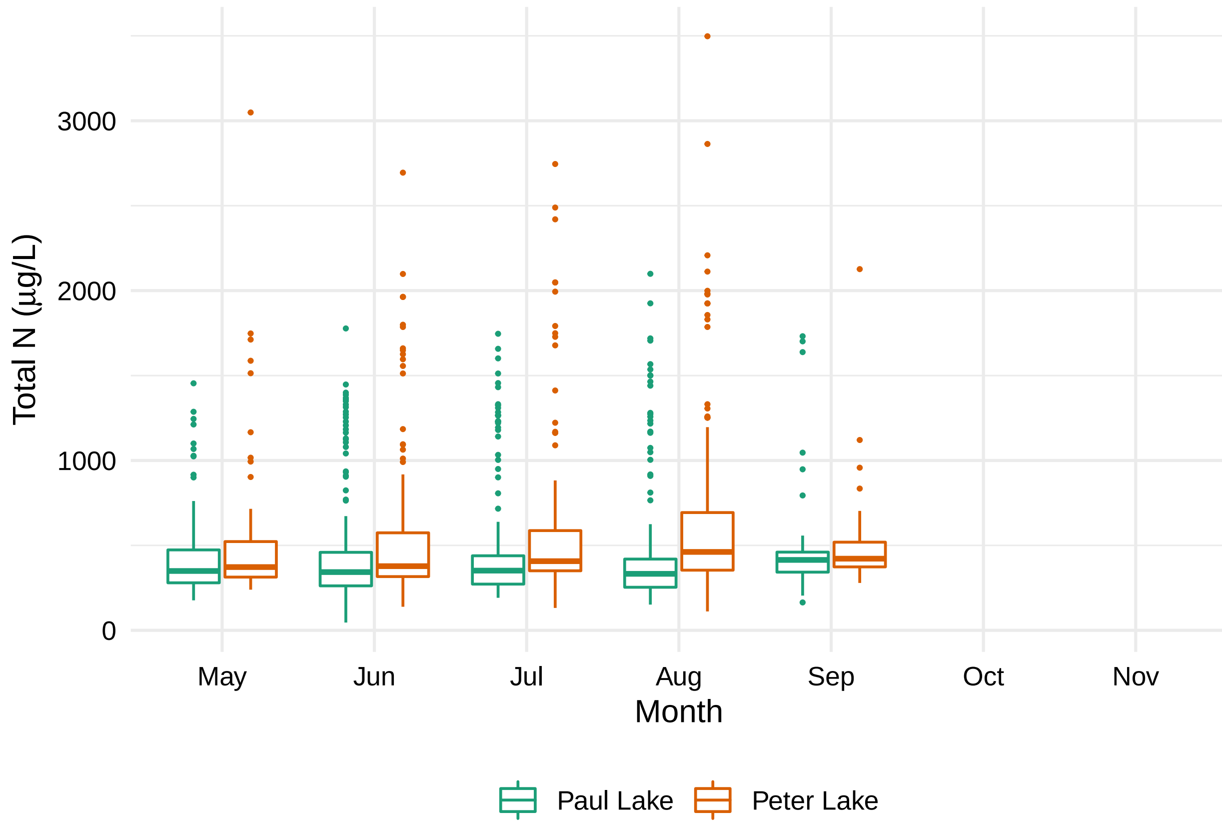
```
## Warning: Removed 16 rows containing missing values (stat_boxplot).
```

```
## Warning: Removed 20713 rows containing non-finite values (stat_boxplot).
```



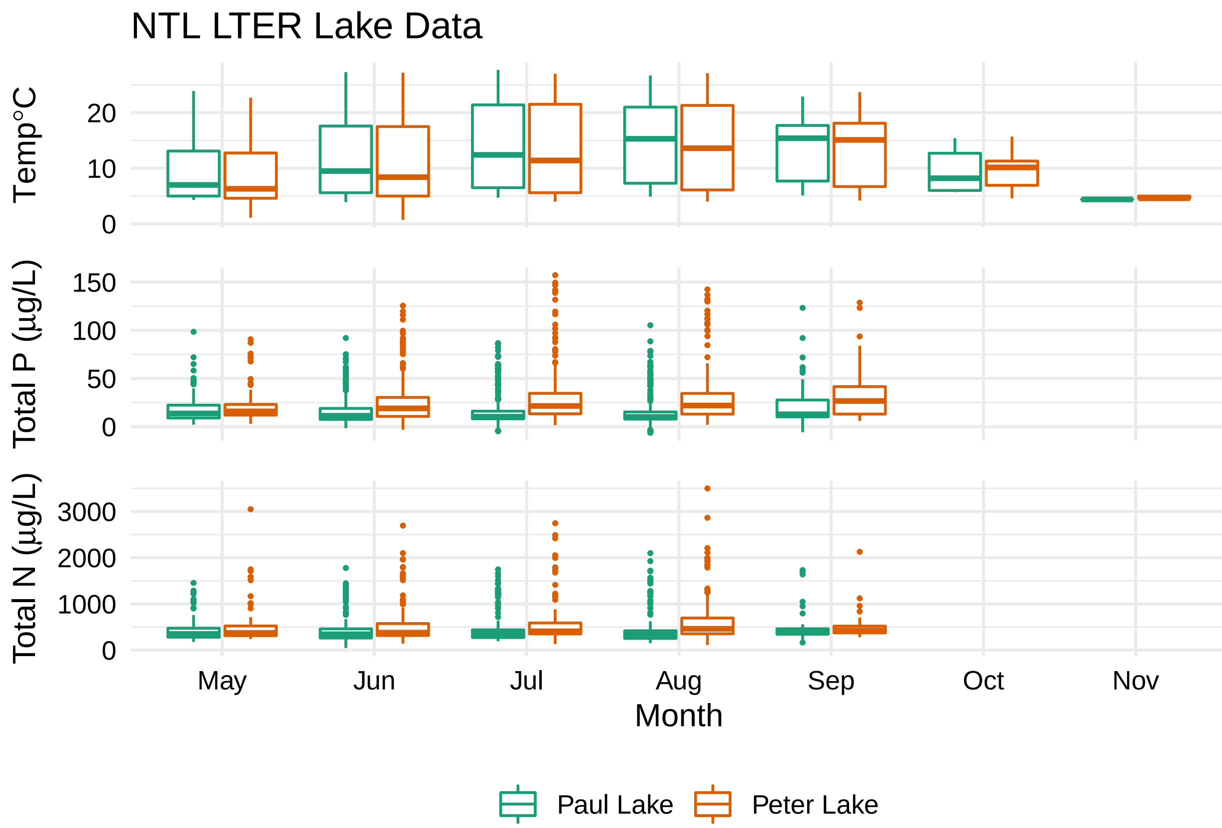

```
boxplotC <- ggplot(PP.nutrients.spread)+
  geom_boxplot(aes(x = month, y = tn_ug, color = lakename),outlier.size = 0.4)+
  labs(x = "Month", y = expression(paste("Total N (", mu, "g/L)")),
       color = "")+
  scale_x_discrete(limits = month.abb[5:11])+
  scale_color_brewer(type = "qual", palette = 2)
print(boxplotC)
```

```
## Warning: Removed 16 rows containing missing values (stat_boxplot).

## Warning: Removed 21567 rows containing non-finite values (stat_boxplot).
```




```r
boxplotA_grid <- boxplotA +
  theme(legend.position = "none",
        axis.title.x = element_blank(),
        axis.text.x = element_blank())+
  labs(title = "NTL LTER Lake Data")


boxplotB_grid <- boxplotB +
  theme(legend.position = "none",
        axis.title.x = element_blank(),
        axis.text.x = element_blank())

plot_grid(boxplotA_grid, boxplotB_grid, boxplotC, nrow = 3, align = 'hv',
          rel_heights = c(1.2, 1,1.85))
```

```
## Warning: Removed 16 rows containing missing values (stat_boxplot).

## Warning: Removed 3550 rows containing non-finite values (stat_boxplot).

## Warning: Removed 16 rows containing missing values (stat_boxplot).

## Warning: Removed 20713 rows containing non-finite values (stat_boxplot).

## Warning: Removed 16 rows containing missing values (stat_boxplot).

## Warning: Removed 21567 rows containing non-finite values (stat_boxplot).

## Warning: Graphs cannot be horizontally aligned unless the axis parameter is set.
## Placing graphs unaligned.
```



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

> Answer: * temperatures are generally higher but also have more variation in the summer months (July - Sep), and cooler with less variation in the spring and fall. Temps are pretty similar between the two lakes * Total P - no obvious seasonal trend, maybe a little higher in summr months, but in general Peter Lake may have higher median total P than Paul in peak summer months. Data are positively skewed. Also, I am not sure why there are negative values for TP concentration…that does not make sense, and there is not explanation in the metadata. * total N - again, no obvious seasonal variation, lakes are pretty similar but Peter lake may have a slightly higher median total N than Paul, especially in mid summer. Data are positively skewed.

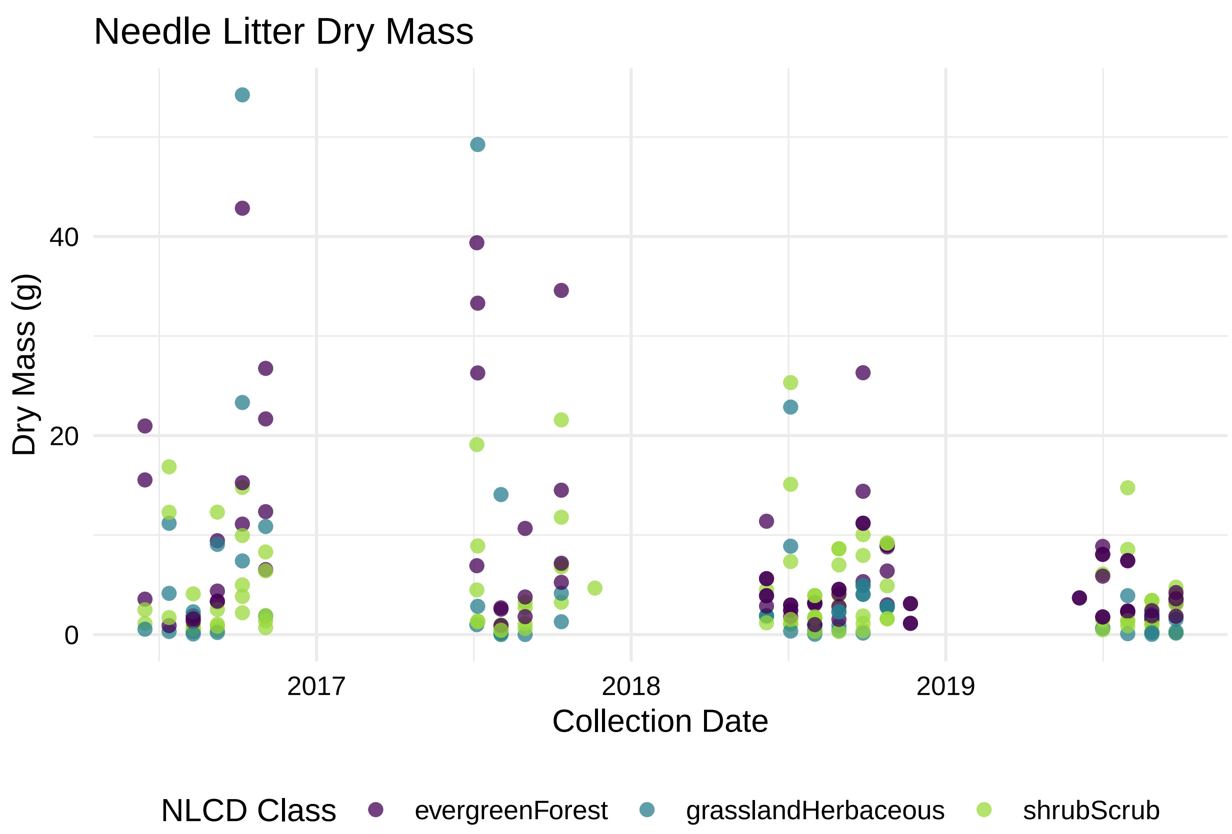
6. [Niwot Ridge] Plot a subset of the litter dataset by displaying only the "Needles" functional group. Plot the dry mass of needle litter by date and separate by NLCD class with a color aesthetic. (no need to adjust the name of each land use)
7. [Niwot Ridge] Now, plot the same plot but with NLCD classes separated into three facets rather than separated by color.

```r
# 6
levels(Litter$functionalGroup)
```

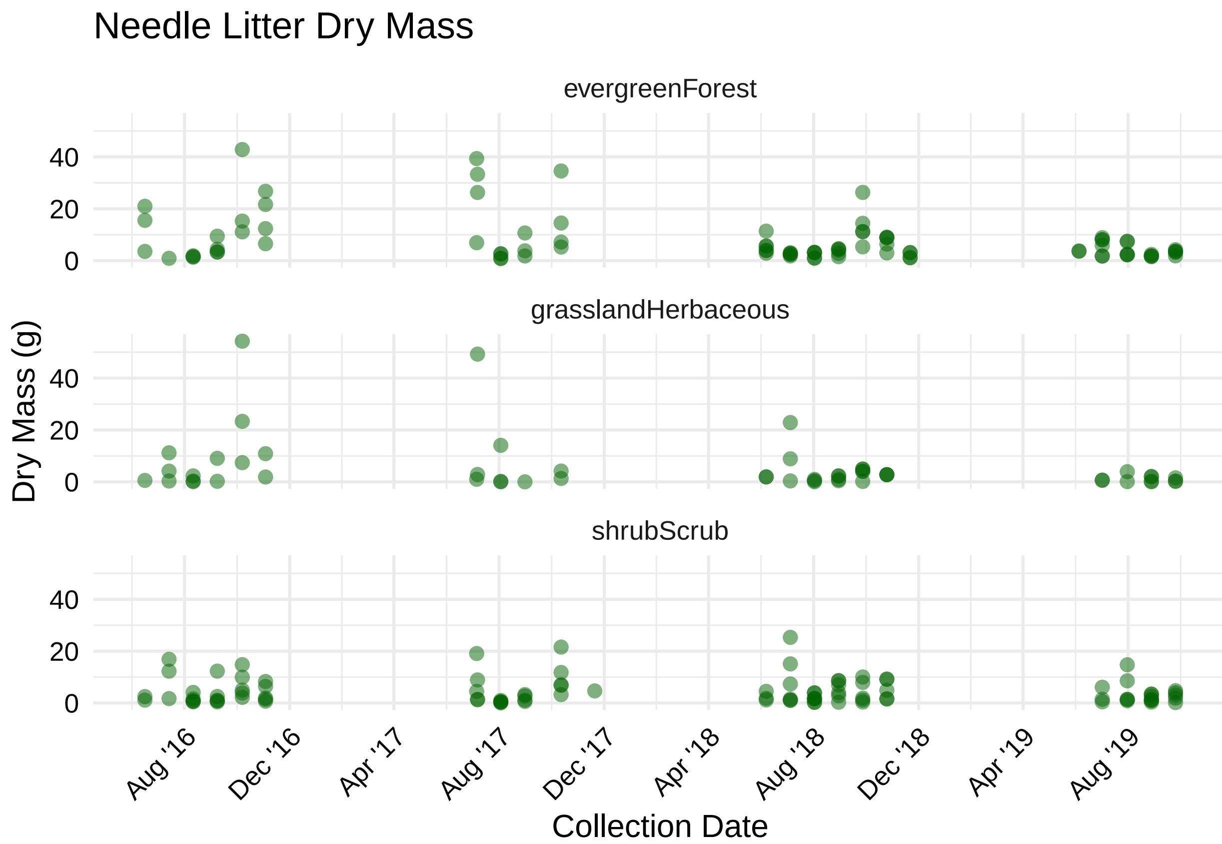
```
## [1] "Flowers"         "Leaves"          "Mixed"           "Needles"
## [5] "Other"           "Seeds"           "Twigs/branches" "Woody material"
```

```r
Litter.needles <- Litter%>%
  filter(functionalGroup == "Needles")

needlesplot1 <- ggplot(Litter.needles)+
  geom_point(aes(x = as.Date(collectDate), y = dryMass,
```

```r
                 color = nlcdClass), size = 2,alpha = 0.75)+
  labs(x= "Collection Date", y = "Dry Mass (g)", color = "NLCD Class", title = "Needle Litter Dry Mass")
  scale_color_viridis_d(option = "viridis", end = .85)

print(needlesplot1)
```




```r
#7
needlesplot_facets <- ggplot(Litter.needles)+
  geom_point(aes(x = as.Date(collectDate), y = dryMass), size = 2,alpha = 0.5, color = "darkgreen")+
  labs(x= "Collection Date", y = "Dry Mass (g)", title = "Needle Litter Dry Mass")+
  scale_color_viridis_d(option = "viridis", end = .85)+
  theme(legend.position = "none")+
  facet_wrap(Litter.needles$nlcdClass, nrow=3, strip.position = "top")+
  scale_x_date(date_breaks = "4 months", date_labels = "%b '%y")+
  theme(axis.text.x = element_text(angle = 45,  hjust = 1))

print(needlesplot_facets)
```



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

Answer: I think plot 7 is more effective at showing the differences between the NLCD classes - plot 6 has many overlapping points and colors and it is difficult to focus on one color at a time whereas plot 7 allows us to quickly compare seasonal trends within and between groups and easily see differences between NLCD classes.