Assignment 5: Data Visualization

Jingze Dai

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 setting up

# loading libraries
library(tidyverse); library(lubridate); library(here); library(cowplot)
```

```
## -- 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 1.9.3     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
## here() starts at /home/guest/ENV872/EDE_Fall2024
##
##
## Attaching package: 'cowplot'
##
##
## The following object is masked from 'package:lubridate':
##
##     stamp
```

```
# verifying home directory
here()
```

```
## [1] "/home/guest/ENV872/EDE_Fall2024"
```

```
# loading datasets
peterPaul.chem.nutrients <-
  read.csv(
    here(
      "Data/Processed_KEY/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Processed.csv"
      ), stringsAsFactors = T)

litter.mass.trap <-
  read.csv(here("Data/Processed_KEY/NEON_NIWO_Litter_mass_trap_Processed.csv"),
           stringsAsFactors = T)

#2 changing dates to date format

# checking data types
glimpse(peterPaul.chem.nutrients)
```

```
## 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~
```

```
glimpse(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~
```

```
# we can see that the collection date for both datasets are factors

# converting dates to date format
peterPaul.chem.nutrients$sampledate <- ymd(peterPaul.chem.nutrients$sampledate)
litter.mass.trap$collectDate <- ymd(litter.mass.trap$collectDate)
```

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 building a theme

mytheme <- theme_classic(base_size = 10) +
  theme(
    # plot background
    plot.background = element_rect(fill = "lightgrey", color = NA),
    # plot title
    plot.title = element_text(size = 8, face = "bold", hjust = 0.5),
    # axes labels
    axis.title.x = element_text(size = 10, face = "bold", color = "black"),
    axis.title.y = element_text(size = 10, face = "bold", color = "black"),
    # axis ticks and lines
    axis.ticks = element_line(color = "black"),
    axis.line = element_line(color = "black"),
    # gridlines
    panel.grid.major = element_line(color = "grey90", size = 0.5),
    # legend
```

```
    legend.position = "right",
    legend.box.size = unit(1, "cm")
  )
```

```
## Warning: The ‘size‘ argument of ‘element_line()‘ is deprecated as of ggplot2 3.4.0.
## i Please use the ‘linewidth‘ argument instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

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()`).

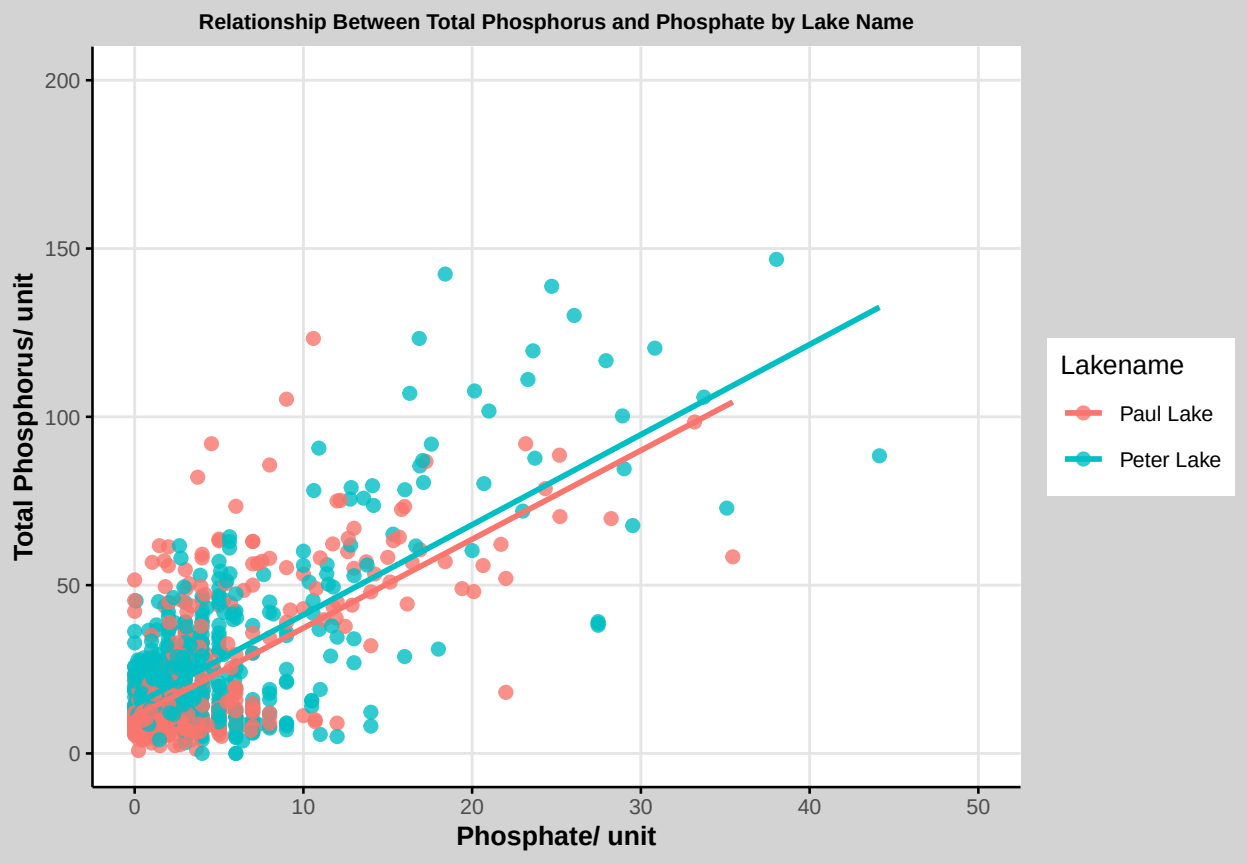
```
#4 Scatter plot with best fit line

# plotting total phosphorus by phosphate, separated by lake
ggplot(peterPaul.chem.nutrients, aes(x = po4, y = tp_ug, color = lakename)) +
  geom_point(size = 2, alpha = 0.8) +
  # best fit line
  geom_smooth(method = "lm", se = FALSE) +
  # hide extreme values
  xlim(0, 50) +
  ylim(0, 200) +
  labs(
    title = "Relationship Between Total Phosphorus and Phosphate by Lake Name",
    x = "Phosphate/ unit",
    y = "Total Phosphorus/ unit",
    color = "Lakename"
  ) +
  mytheme
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```

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

```
## Warning in plot_theme(plot): The ‘legend.box.size‘ theme element is not defined
## in the element hierarchy.
```

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

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 box plots
library(grid)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
# (a) Boxplot for temperature by month
temp_plot <- ggplot(peterPaul.chem.nutrients,
                    aes(x = month, y = temperature_C, fill = lakename)) +
  geom_boxplot() +
```

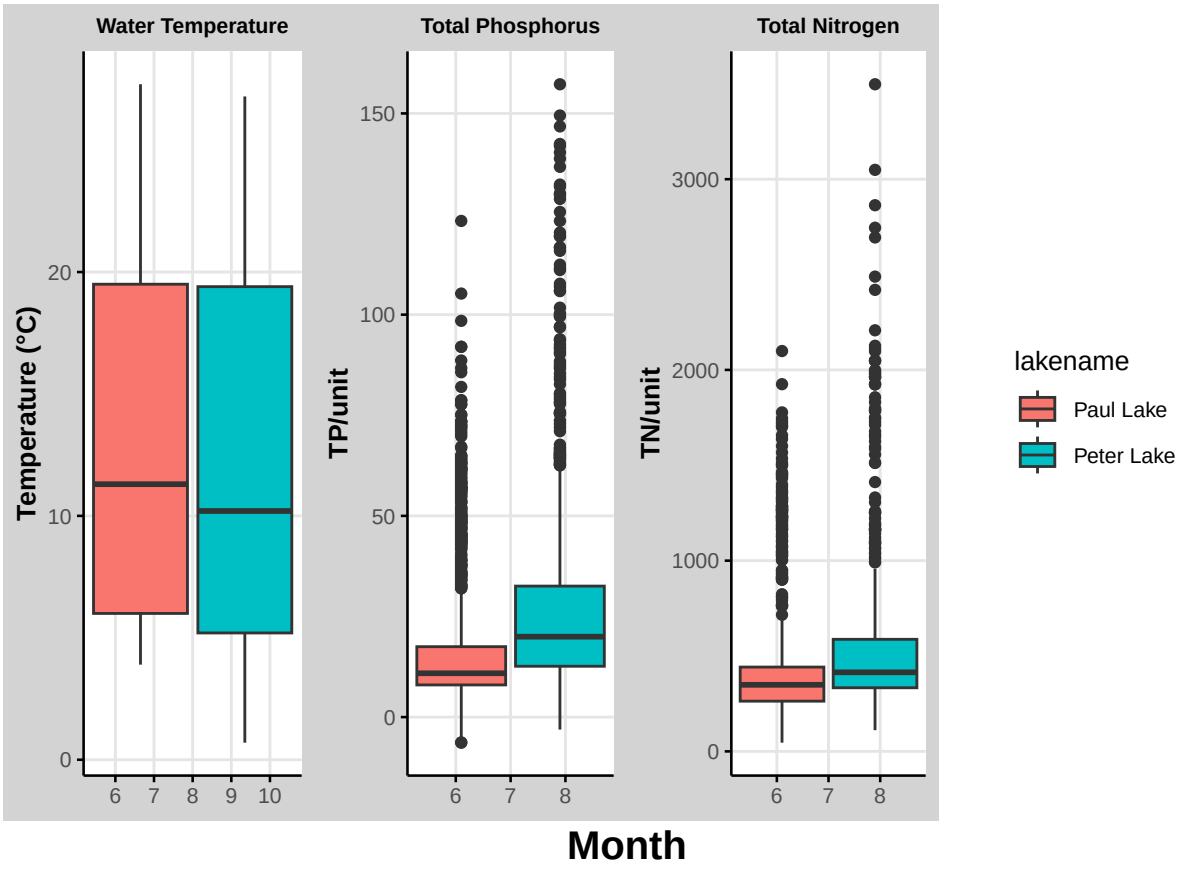
```
  labs(title = "Water Temperature", y = "Temperature (°C)") +
  mytheme +
  theme(legend.position = "none", axis.title.x = element_blank())

# (b) Boxplot for total phosphorus by month
tp_plot <- ggplot(peterPaul.chem.nutrients,
                  aes(x = month, y = tp_ug, fill = lakename)) +
  geom_boxplot() +
  labs(title = "Total Phosphorus", y = "TP/unit") +
  mytheme +
  theme(legend.position = "none", axis.title.x = element_blank())

# (c) Boxplot for total nitrogen by month
tn_plot <- ggplot(peterPaul.chem.nutrients,
                  aes(x = month, y = tn_ug, fill = lakename)) +
  geom_boxplot() +
  labs(title = "Total Nitrogen", y = "TN/unit") +
  mytheme +
  theme(legend.position = "none", axis.title.x = element_blank())

legend <- get_legend(tn_plot + theme(legend.position = "right"))
```

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

## Warning in plot_theme(plot): The `legend.box.size` theme element is not defined
## in the element hierarchy.

## Warning in get_plot_component(plot, "guide-box"): Multiple components found;
## returning the first one. To return all, use `return_all = TRUE`.
```

```
# combining plots
combined_plot <- plot_grid(temp_plot, tp_plot, tn_plot, legend, ncol = 4)
```

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

## Warning in plot_theme(plot): The `legend.box.size` theme element is not defined
## in the element hierarchy.

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

## Warning in plot_theme(plot): The `legend.box.size` theme element is not defined
## in the element hierarchy.

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

## Warning in plot_theme(plot): The `legend.box.size` theme element is not defined
## in the element hierarchy.
```

```
# adding common x-axis
x.grob <- textGrob("Month",
                   gp=gpar(fontface="bold", fontsize=15))
grid.arrange(arrangeGrob(combined_plot, bottom = x.grob))
```



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

Answer: The variable across lakes are not very different in terms of temperature. But the phosphorus and nitrogen content is higher in peter

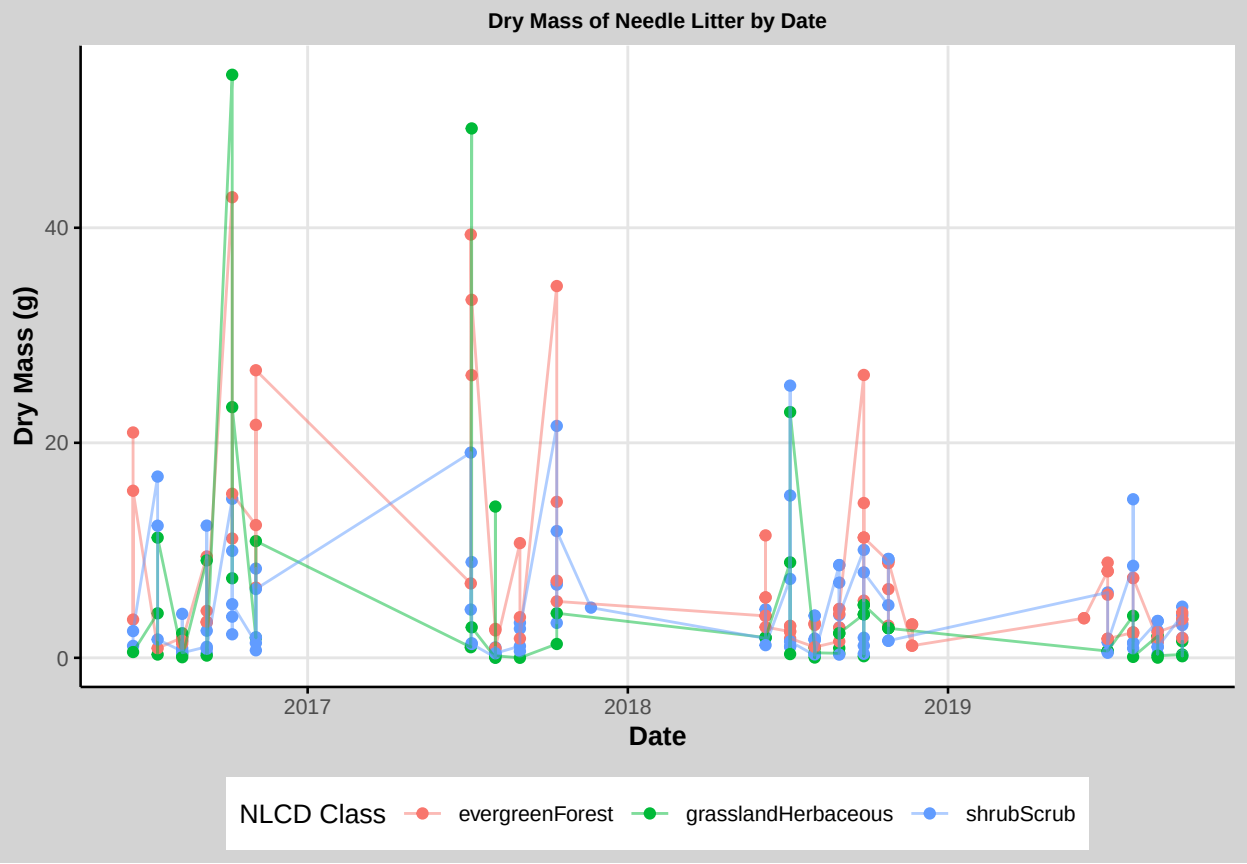
6. [Niwot Ridge] Plot a subset of the litter dataset by displaying only the "Needles" functional group. Plot the dry mass of needle litter by date and separate by NLCD class with a color aesthetic. (no need to adjust the name of each land use)

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

```
#6 Needle plot

# filtering out data
needles_data <- litter.mass.trap %>%
  filter(functionalGroup == "Needles")

# creating needle plot
needle_plot <- ggplot(needles_data,
```

```
                      aes(x = collectDate, y = dryMass, color = nlcdClass)) +
  geom_point() +
  geom_line(aes(group = nlcdClass), alpha = 0.5) +
  labs(title = "Dry Mass of Needle Litter by Date",
       x = "Date",
       y = "Dry Mass (g)",
       color = "NLCD Class") +
  mytheme +
  theme(legend.position = "bottom")  # move legend to the bottom

needle_plot
```

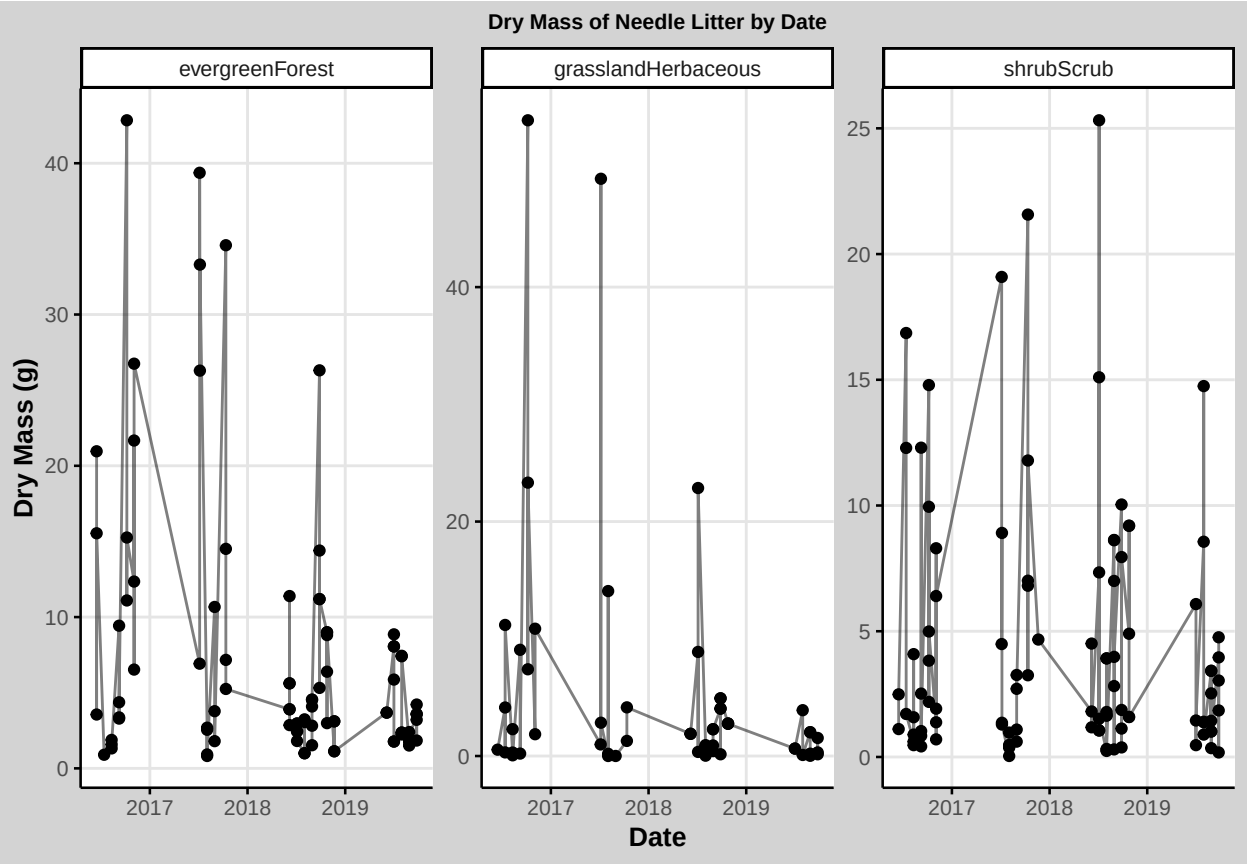
```
## Warning in plot_theme(plot): The 'legend.box.size' theme element is not defined
## in the element hierarchy.
```




```
#7 creating 3 facets

# creating faceted plot
needle_plot_facet <- ggplot(needles_data,
                      aes(x = collectDate, y = dryMass)) +
  geom_point() +
  geom_line(aes(group = nlcdClass), alpha = 0.5) +
  labs(title = "Dry Mass of Needle Litter by Date",
       x = "Date",
```

```
       y = "Dry Mass (g)") +
  mytheme +
  theme(legend.position = "none") +
  facet_wrap(~ nlcdClass, scales = "free_y")

needle_plot_facet
```

```
## Warning in plot_theme(plot): The ‘legend.box.size‘ theme element is not defined
## in the element hierarchy.
```



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

Answer: The plot in Q6 is more effective, because it can compare the different facets more directly by observing the patterns of different colors. It allows a direct comparison and is better due to its simplicity.