Assignment 3: Data Exploration

Reed Leon-Hinton

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Exploration.

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_A03_DataExploration.Rmd") prior to submission.

The completed exercise is due on <>.

Set up your R session

1. Check your working directory, load necessary packages (tidyverse), and upload two datasets: the ECOTOX neonicotinoid dataset (ECOTOX_Neonicotinoids_Insects_raw.csv) and the Niwot Ridge NEON dataset for litter and woody debris (NEON_NIWO_Litter_massdata_2018-08_raw.csv). Name these datasets "Neonics" and "Litter", respectively.

```
# clearing the environment (It's a pet peeve)
remove(list = ls())

# checking the working directory
getwd()
```

```
## [1] "C:/Users/shado/Documents/Graduate School Stuff/ENVIRON 872 - Environmental Data Analytics/Envir
```

```
# installing and loading the tidyverse package
# install.packages(tidyverse)
library(tidyverse)

# creating the ECOTOX neonicotinoid dataset
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv")

# creating the Niwot Ridge NEON dataset
Litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv")
```

Learn about your system

2. The neonicotinoid dataset was collected from the Environmental Protection Agency's ECOTOX Knowledgebase, a database for ecotoxicology research. Neonicotinoids are a class of insecticides used widely in agriculture. The dataset that has been pulled includes all studies published on insects. Why

might we be interested in the ecotoxicologoy of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

Answer: The ecotoxicology of insects has the potential to affect the natural environment. If there are toxicants introduced that have an effect on animals higher up the food chain, the introduction of a single chemical could have a substantial and disastrous effect on a much wider scale. Additionally, you should know which insects are affected by the new insecticide to ensure you are only killing the desired targets and not killing all insects in the area.

3. The Niwot Ridge litter and woody debris dataset was collected from the National Ecological Observatory Network, which collectively includes 81 aquatic and terrestrial sites across 20 ecoclimatic domains. 32 of these sites sample forest litter and woody debris, and we will focus on the Niwot Ridge long-term ecological research (LTER) station in Colorado. Why might we be interested in studying litter and woody debris that falls to the ground in forests? Feel free to do a brief internet search if you feel you need more background information.

   Answer: The type of litter and woody debris on the forest floor will have a substantial influence on the insects found in the area. Knowing the potential insect inhabitants of the area, through comparison between other areas with similar conditions, may give insight into the total effects of the insecticide.

4. How is litter and woody debris sampled as part of the NEON network? Read the NEON_Litterfall_UserGuide.pdf document to learn more. List three pieces of salient information about the sampling methods here:

   Answer: * 1.) They conduct the analysis on wood debris which dropped from the forest canopy and captured in ground traps with a diameter of between 2 and 50 cm. * 2.) They only sample areas which contain woody vegetation that is greater than 2 meters tall. * 3.) Ground traps are sampled once each year. Elevated traps sampling frequency varies depending on the vegetation present at the site.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
# getting the dimensions of the Neonics dataset and storing them for later use
Neonics_len <- dim(Neonics)[1]
Neonics_wid <- dim(Neonics)[2]

# printing the output of the dimensions
print(paste("The Neonics dataset has", Neonics_len, "recorded values for", Neonics_wid, "variables."))
```

```
## [1] "The Neonics dataset has 4623 recorded values for 30 variables."
```

6. Using the `summary` function on the "Effects" column, determine the most common effects that are studied. Why might these effects specifically be of interest?

```
# utilizing summary on the effects column
summary(Neonics$Effect)
```

```
##    Length     Class      Mode
##      4623 character character
```

```
# the previous method did not work. trying another idea by separating the column
# from the data set, changing the class to a factor variable,
# and then running the summary on it.
Neonics_effect <- as.factor(Neonics$Effect)

summary(Neonics_effect) # this worked.
```

```
##     Accumulation        Avoidance         Behavior     Biochemistry
##               12              102              360               11
##          Cell(s)      Development        Enzyme(s) Feeding behavior
##                9              136               62              255
##         Genetics           Growth        Histology       Hormone(s)
##               82               38                5                1
##    Immunological     Intoxication       Morphology        Mortality
##               16               12               22             1493
##       Physiology       Population     Reproduction
##                7             1803              197
```

Answer: The most common effects are effects on the population of a species and mortality. These two effects directly impact the potential consequences discussed earlier: potential fatalities of individuals not in the desired population. However, mortality and population also represent success of an insecticide if they are members of the target population.

7. Using the `summary` function, determine the six most commonly studied species in the dataset (common name). What do these species have in common, and why might they be of interest over other insects? Feel free to do a brief internet search for more information if needed.

```
# utilizing the same factor method described above combined with a sort and head operation.
Neonics_summary <- summary(as.factor(Neonics$Species.Common.Name))

# removing the other category from the summary results to make the top species name sort
# work better
Neonics_summary <- sort((Neonics_summary), decreasing = TRUE)

# Now that it is sorted, we remove the first element
Neonics_summary <- Neonics_summary[-1]

# Now we establish the top six species studied and print the results
top_species <- head(Neonics_summary, 6)
top_species
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee      Italian Honeybee
##                   152                   140                   113
```

Answer: All of these species most frequently studied are insects which build hives and are also having population sustainability issues. They are endangered of their populations reaching critically low levels and, thus, this insecticide having an effect on their population is detrimental.

8. Concentrations are always a numeric value. What is the class of Conc.1..Author. in the dataset, and why is it not numeric?

```
# looking into the class for the Conc.1..Author variable
class(Neonics$Conc.1..Author.)
```

```
## [1] "character"
```

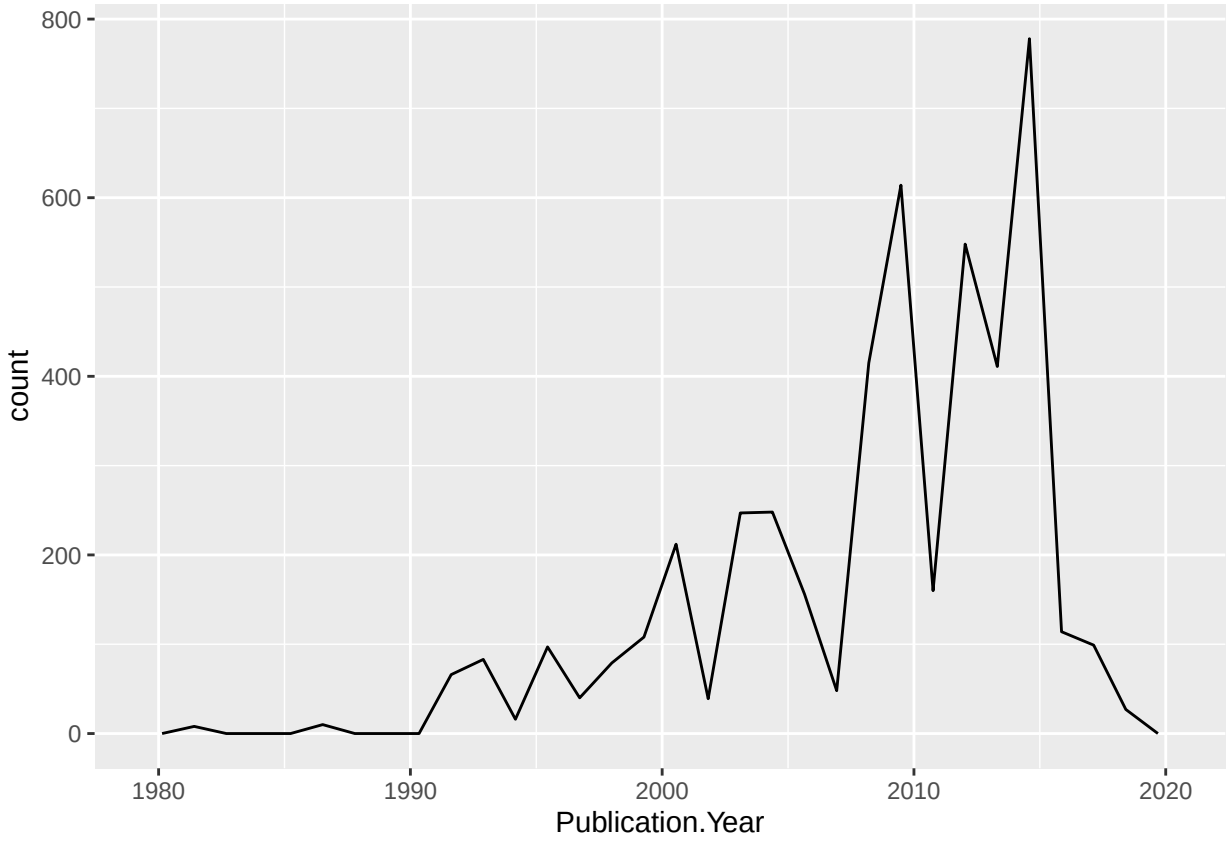
Answer: The Conc.1..Author variable is non-numeric and, instead, is a character class. This is because the data in the column was not cleanly formatted when reported. There are some "/" and "NR" values within the column which caused R to import them as a character class rather than a numeric class.

Explore your data graphically (Neonics)

9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
# writing the graphing function, here we go...
ggplot(data = Neonics, aes(Publication.Year)) +
  geom_freqpoly()
```

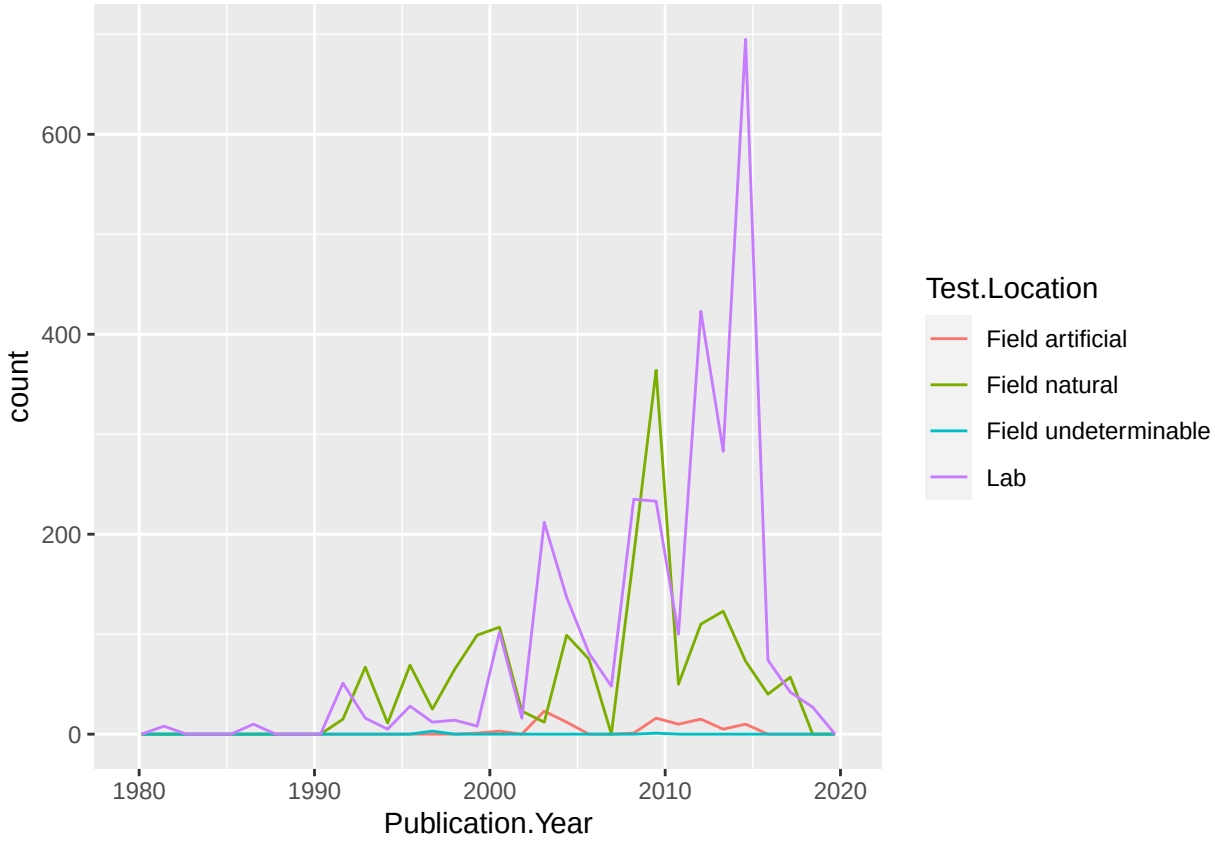
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
# that last graph needed work, adding different lines based on test location.
ggplot(data = Neonics, aes(Publication.Year, colour = Test.Location)) +
  geom_freqpoly()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

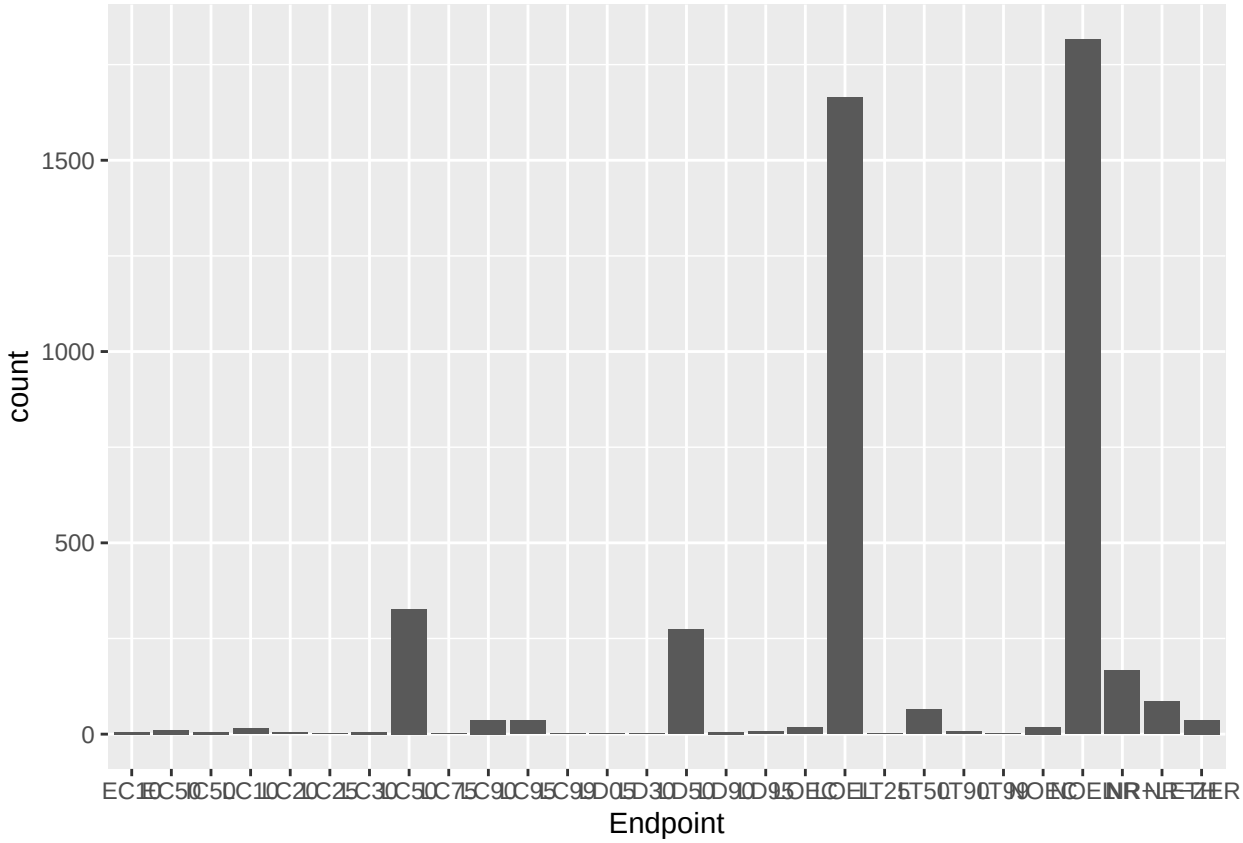


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are the lab and the natural field. They are seemingly inversely correlated with one another, which would make sense. If natural conditions are not conducive to running a consistent test in a natural environment, reproducing the conditions in a lab or artificial field will be necessary. The most common testing location does vary over time and is likely related to this natural conditions consideration.

11. Create a bar graph of Endpoint counts. What are the two most common end points, and how are they defined? Consult the ECOTOX_CodeAppendix for more information.

```
# making the endpoint a factor variable
Neonics$Endpoint <- as.factor(Neonics$Endpoint)

# making a bar graph of the Endpoint counts
ggplot(data = Neonics, aes(Endpoint)) +
  geom_bar()
```




```
# using the method described above to find the two most common
endpoint_summary <- summary(Neonics$Endpoint)
endpoint_summary <- sort(endpoint_summary, decreasing = TRUE)
head(endpoint_summary, 2) # I use head because it feels more legible later,
```

```
## NOEL LOEL
## 1816 1664
```

```
# I realize you could also use endpoint_summary[1:2] for the same result.
```

Answer: The two most common endpoints are NOEL and LOEL. NOEL indicates that there was no-observable-effect-level where the highest dose produced effects not significantly different from the control group. Similarly in naming convention, LOEL indicates the lowest-observable-effect-level which is the lowest dose produced effects that were significantly different than the control group.

Explore your data (Litter)

12. Determine the class of collectDate. Is it a date? If not, change to a date and confirm the new class of the variable. Using the `unique` function, determine which dates litter was sampled in August 2018.

```
# will be using lubridate for the date manipulation
# install.packages("lubridate")
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
```

```
##
##     date, intersect, setdiff, union
```

```
# looking at the class of collectDate in Litter
class(Litter$collectDate) #it is a character value.
```

```
## [1] "character"
```

```
# converting the value to a date
Litter$collectDate <- ymd(Litter$collectDate)

# checking to ensure it was formatted correctly.
class(Litter$collectDate)
```

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

```
# Litter$collectDate # this is commented out due to not wanting it printed in the final report
# it is very long.

# learning and using the unique function
collected_dates <- unique(Litter$collectDate)

#Printing the results cleanly.
print(paste("The results were collected on", collected_dates[1], "and",
            collected_dates[2]))
```

```
## [1] "The results were collected on 2018-08-02 and 2018-08-30"
```

13. Using the `unique` function, determine how many plots were sampled at Niwot Ridge. How is the information obtained from `unique` different from that obtained from `summary`?

```
# using the unique function to obtain the overall number of plots sampled.
# making the plot a factor first.
Litter$plotID <- as.factor(Litter$plotID)

unique(Litter$plotID) # Unique gives the actual factors themselves
```

```
##  [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

```
summary(Litter$plotID) # whereas Summary gives the count of each factor
```

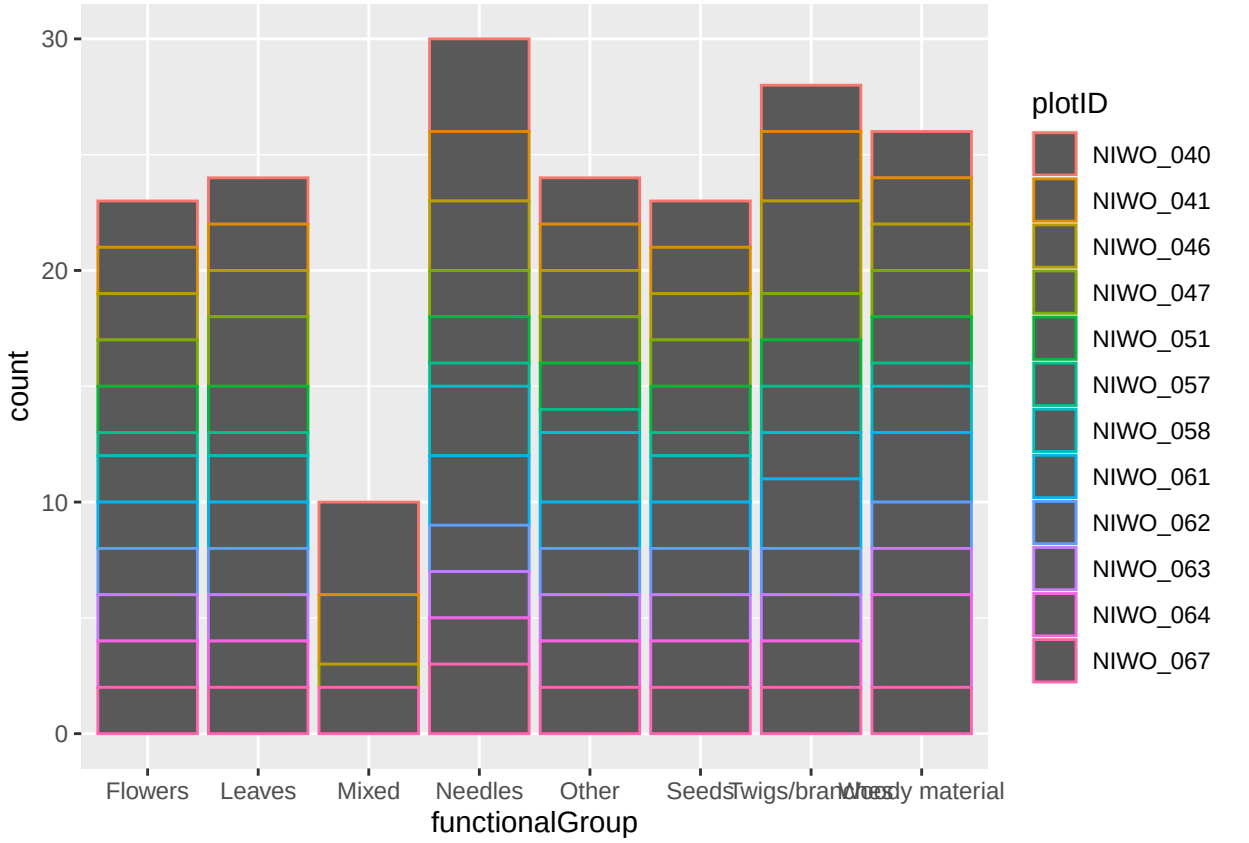
```
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061
##       20       19       18       15       14        8       16       17
## NIWO_062 NIWO_063 NIWO_064 NIWO_067
##       14       14       16       17
```

Answer: Using unique tells you the names of the factor variable as the output, whereas the summary function shows you the names but is primarily used to get the count of each factor.

14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

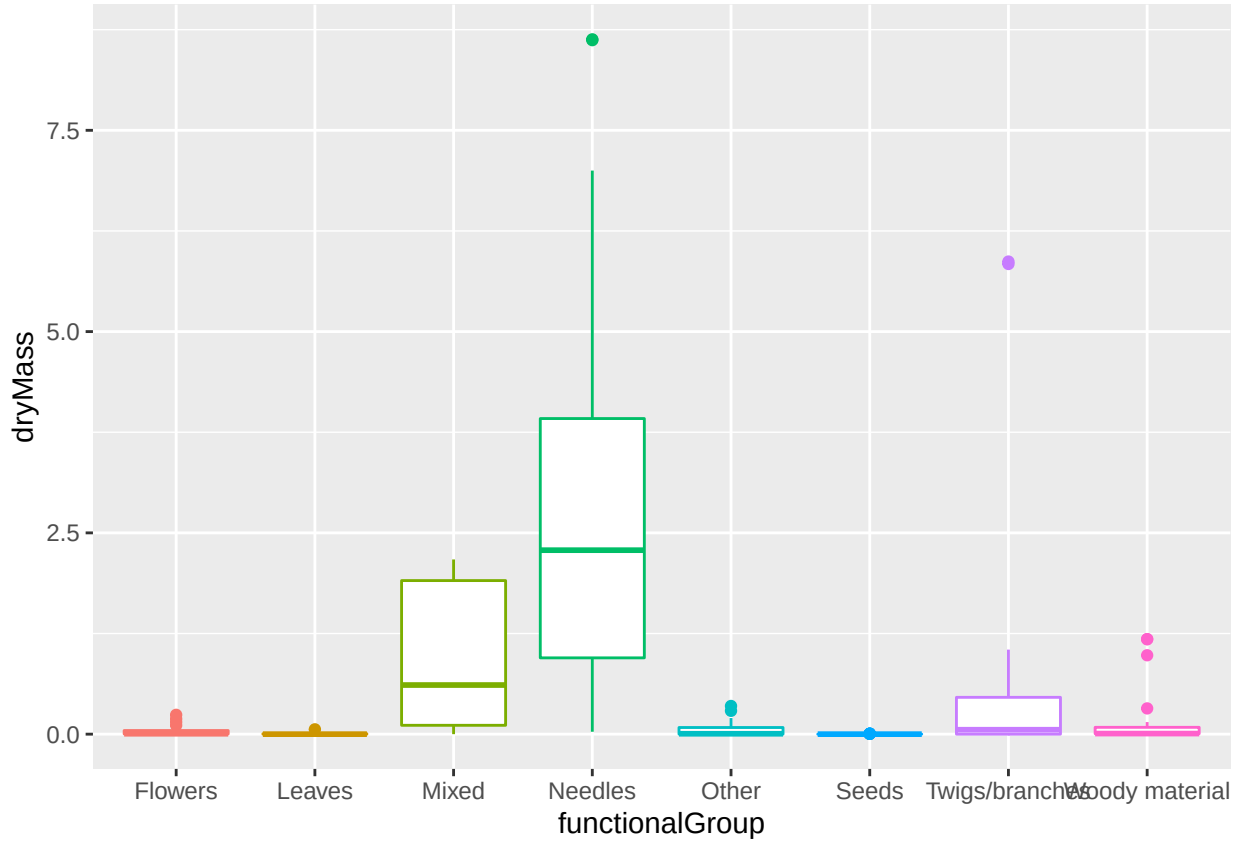
```
# starting by making functionalgroup a factor
Litter$functionalGroup <- as.factor(Litter$functionalGroup)


# creating the bar graph with functionalGroup with stacked bars showing the PlotID.
ggplot(data = Litter, aes(functionalGroup, colour = plotID)) +
```

```
  geom_bar()
```

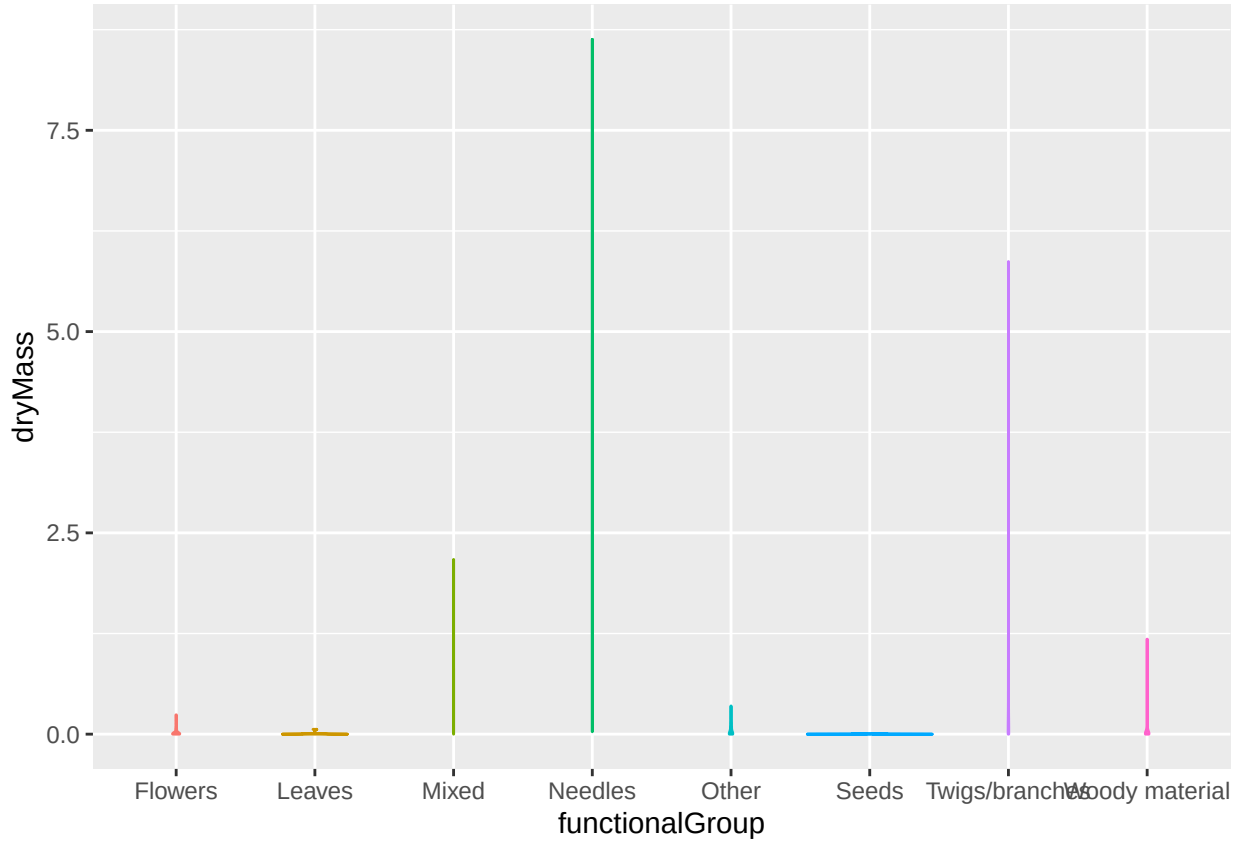


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
# creating a boxplot of the dryMass by functionalgroup
ggplot(data = Litter, aes(x = functionalGroup, y = dryMass, colour = functionalGroup)) +
  geom_boxplot() +
  theme(legend.position = "none")
```




```
# creating a violin plot of dryMass by functionalgroup
ggplot(data = Litter, aes(x = functionalGroup, y = dryMass, colour = functionalGroup)) +
  geom_violin() +
  theme(legend.position = "none")
```




```
# the violin plot looks less than ideal... doing a quick analysis to utilize in my explanation below
summary(Litter$functionalGroup)
```

```
##        Flowers         Leaves          Mixed        Needles          Other
##             23             24             10             30             24
##          Seeds Twigs/branches Woody material
##             23             28             26
```

```
# trying to use the tapply to get summary statistics by factor variable
functional_medians <- tapply(Litter$dryMass, Litter$functionalGroup, median)
functional_medians
```

```
##        Flowers         Leaves          Mixed        Needles          Other
##         0.0050         0.0000         0.6100         2.2850         0.0050
##          Seeds Twigs/branches Woody material
##         0.0000         0.0550         0.0075
```

```
# I liked those results, now sorting by highest median mass
functional_medians <- sort(functional_medians, decreasing = TRUE)
functional_medians
```

```
##        Needles          Mixed Twigs/branches Woody material        Flowers
##         2.2850         0.6100         0.0550         0.0075         0.0050
##          Other         Leaves          Seeds
##         0.0050         0.0000         0.0000
```

```
# doing the same thing for means
functional_means <- tapply(Litter$dryMass, Litter$functionalGroup, mean)
functional_means
```

```
##        Flowers         Leaves          Mixed        Needles          Other
##    0.039782609    0.003750000    0.965000000    2.705000000    0.058125000
##          Seeds Twigs/branches Woody material
##    0.001086957    0.623035714    0.165192308
```

```
functional_means <- sort(functional_means, decreasing = TRUE)
functional_means
```

```
##        Needles          Mixed Twigs/branches Woody material          Other
##    2.705000000    0.965000000    0.623035714    0.165192308    0.058125000
##        Flowers         Leaves          Seeds
##    0.039782609    0.003750000    0.001086957
```

Why is the boxplot a more effective visualization option than the violin plot in this case?

> Answer: There are relatively few observations for each category, with Needles having the most at 30 observations. Therefore, utilization of a Violin plot, which shows frequency of each value occurring in order to gain the width element to each violin, is not advised as it will simply show up as multiple straight lines. However, a box plot, which instead shows the quartiles, median, minimum, and maximum is a much better visualization for smaller sets of data.

What type(s) of litter tend to have the highest biomass at these sites?

> Answer: The three highest biomass types of litter based upon the median values are the:
>
> "Needles" - 2.2850
>
> "Mixed" - 0.6100
>
> "Twigs/branches" - 0.0550
>
> If you look at mean, you see the following:
>
> "Needles" - 2.7050
>
> "Mixed" - 0.9650
>
> "Twigs/branches" - 0.6230