Assignment 3: Data Exploration

Allison Barbaro

Spring 2023

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A03_DataExploration.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. Assign a useful **name to each code chunk** and include ample **comments** with your code.
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.
7. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai.

TIP: If your code extends past the page when knit, tidy your code by manually inserting line breaks.

TIP: If your code fails to knit, check that no `install.packages()` or `View()` commands exist in your code.

Set up your R session

1. Check your working directory, load necessary packages (tidyverse, lubridate), 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. Be sure to include the subcommand to read strings in as factors.

```
getwd()
```

```
## [1] "/home/guest/R/EDA-Spring2023"
```

```
library(tidyverse)
library(lubridate)
library(ggplot2)

Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",stringsAsFactors = TRUE)
Litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv",stringsAsFactors = TRUE)
#this chunk checked our workspace, loaded necessary packages, and uploaded our two datasets.
```

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 ecotoxicology of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: Neonicotinoids are extremely harmful chemicals that kill pests indiscriminantly. They kill important pollinator species like bees, and they can contaminate soils and water, and therefore harm entire ecosystems.

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: Forest litter and debris protect forest soils from drying, extreme temperatures, and erosion. Litter is an important part of forest ecosystems.

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. Ground traps are sampled once per year, while the frequency of sampling for elevated traps varies dependent upon vegetation present at each site. 2. Sampling occurs at sites with woody vegetation that is greater than 2 meters tall. 3. Sampling occurs in tower plots.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

```
## [1] 4623   30
```

```
#this provided the dimensions of our dataset.
```

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

```
summary(Neonics$Effect)
```

```
##     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
```

```
#The most common effects studied are mortality, population, and behavior.
```

Answer: The most common effects of these chemicals on the study insects are of interest because they could represent the most commonly seen impacts of the use of neonics in the real world.

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.[TIP: The `sort()` command can sort the output of the summary command...]

```
summary(Neonics$Species.Common.Name , 6)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee               (Other)
##                   152                   140                  3196
```

```
#this gave the us the six most commonly studied species in our dataset.
```

Answer: The six most commonly studied species in the dataset are the Honey Bee, the Parasitic Wasp, the Buff Tailed Bumblebee, the Carniolan Honey Bee, the Bumble Bee, and Other. These species are all pollinators. These could be of interest over other insects because they are extrmely important to entire ecosystems, and to human agriculture and food systems.

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

```
class('Conc.1..Author')
```

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

```
#this determined the class of the conc.1..author item in our dataset.
```

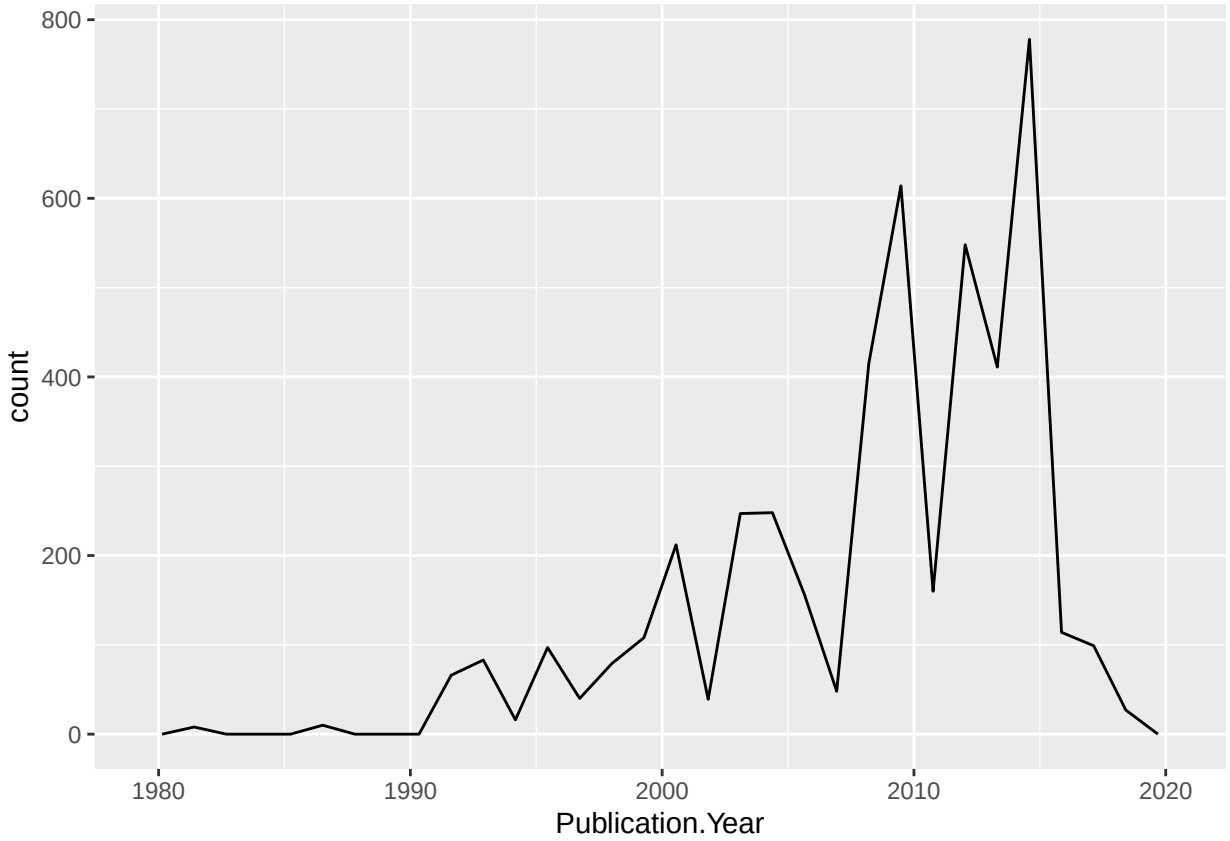
Answer: The class is character. It is not numeric because some of the entries in this column are NR or contain "/".

Explore your data graphically (Neonics)

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

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year))
```

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

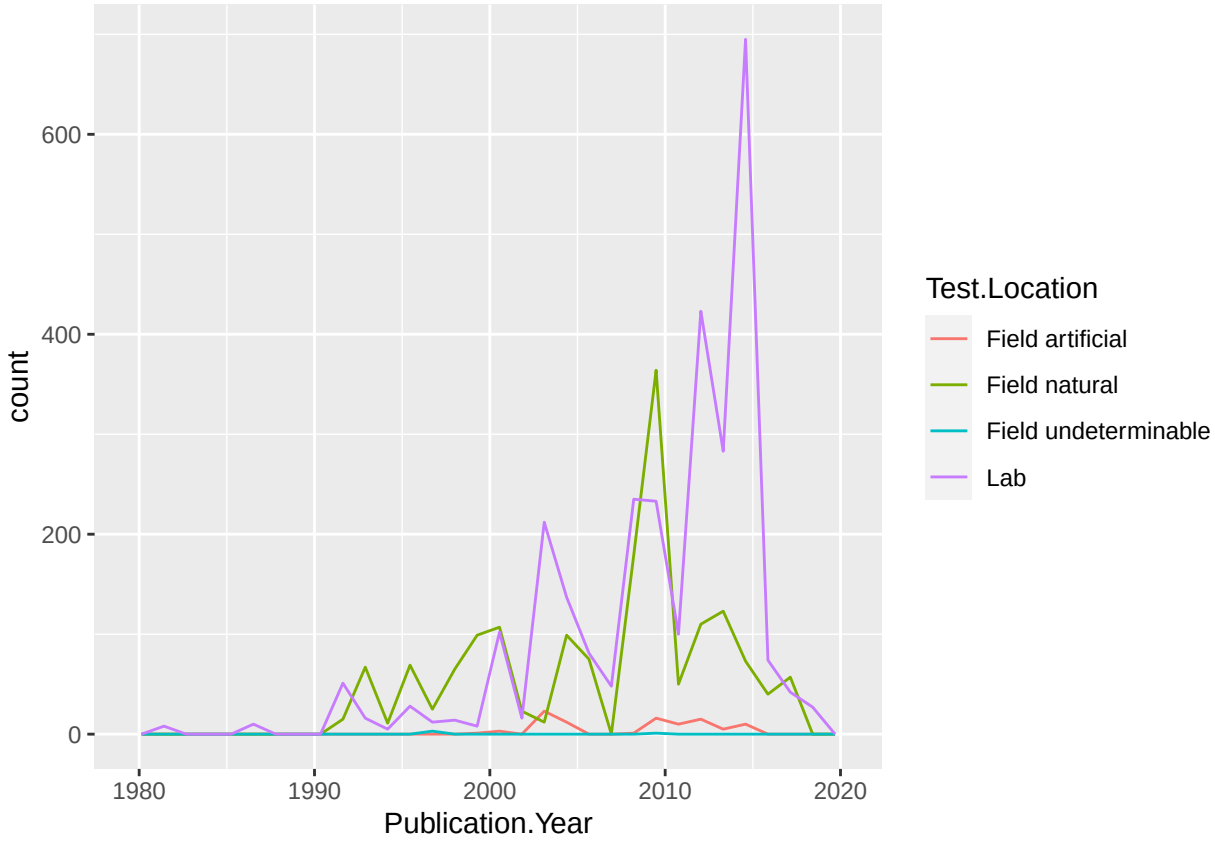



```
#this created a histogram of publication year.
```

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

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location))
```

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




```
#this created a histogram of publication year and test locations.
```

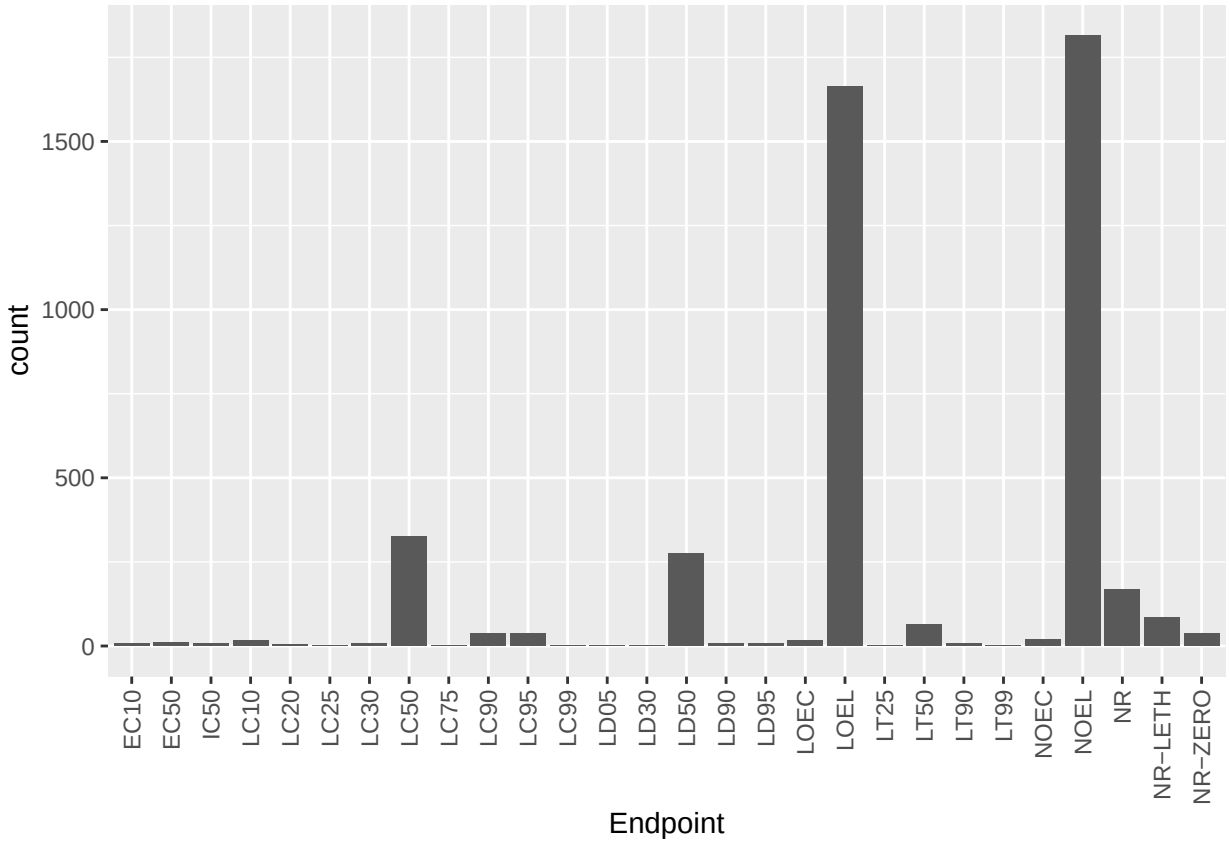
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test location is the Lab, and this does differ slightly over time. At one point between 2000 - 2010, "Field Natural" is the most common test location.

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.

[**TIP**: Add `theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))` to the end of your plot command to rotate and align the X-axis labels...]

```
ggplot(Neonics) +
  geom_bar((aes(x = Endpoint))) +
             theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```




```
#this created a bar graph of endpoint counts.
```

Answer: The two most common endpoints are LOEL and NOEL. LOEL stands for Lowest Observable Effect Level. This is the lowest dose producing effects that were significantly different from the control. NOEL stands for No Observable Effect Level. This is the highest dose producing effects not significantly different from the control.

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.

```
class('collectDate') #it is a character
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate)
unique(Litter$collectDate)
```

```
## [1] "2018-08-02" "2018-08-30"
```

```
#collectDate was turned into a date class, then the dates (written on the comment below) were extracted
# August 2, 2018 and August 30, 2018
```

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`?

```
length(unique(Litter$plotID))
```

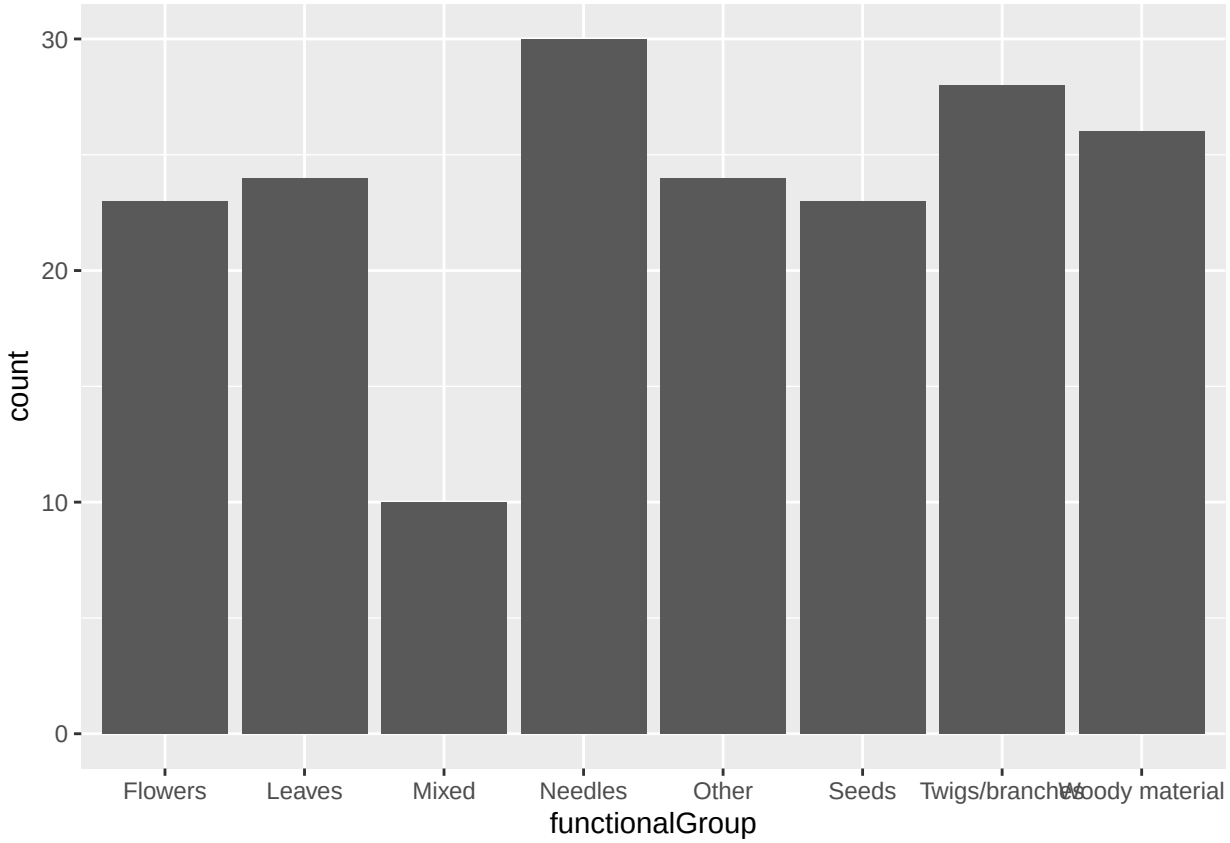
```
## [1] 12
```

```
#this gives us the number of plots samples at Niwot Ridge.
```

Answer: 12 plots were sampled at Niwot Ridge. Unique tells us how many different plotIDs there are, while summary tells us how many of each plotID exist in our dataset.

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.

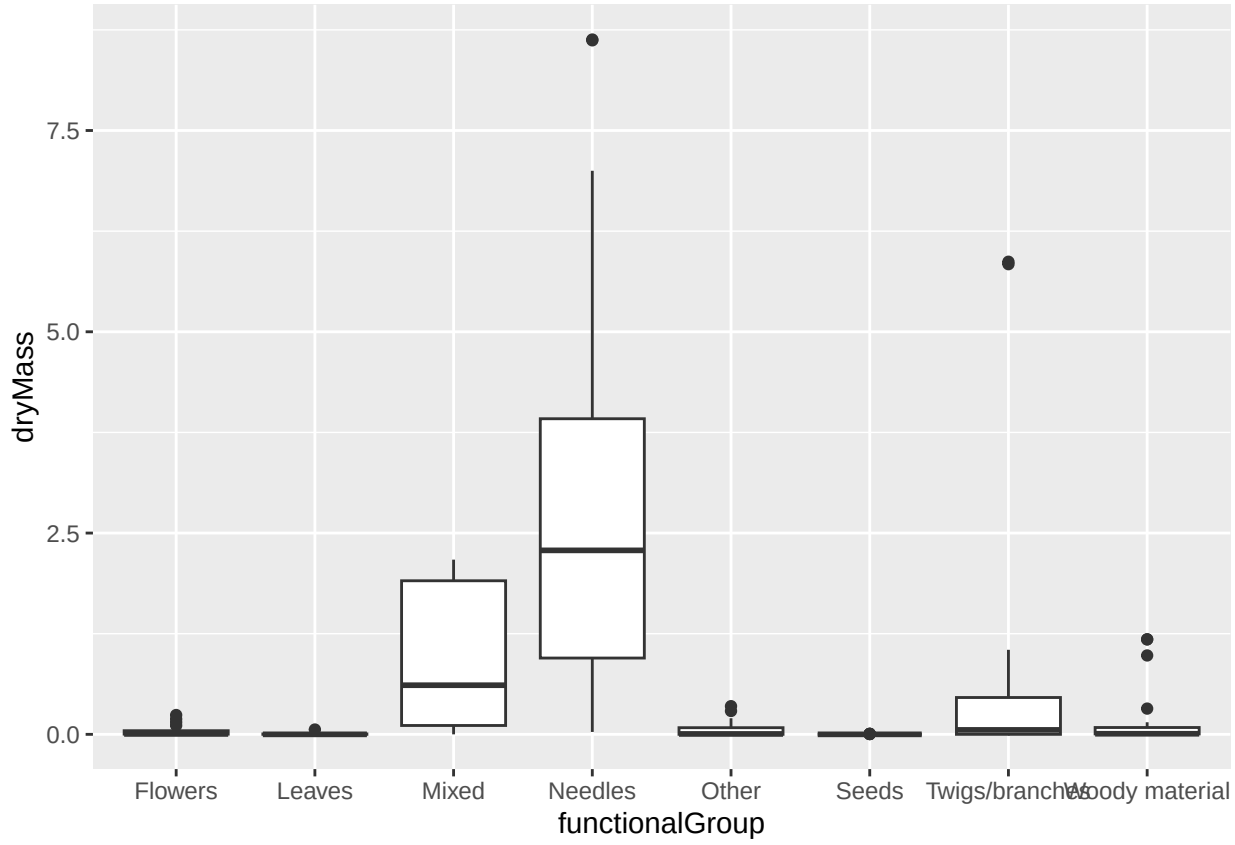
```
ggplot(Litter) +
  geom_bar((aes(x = functionalGroup)))
```



```
#this creates a bar graph of functional group counts.
```

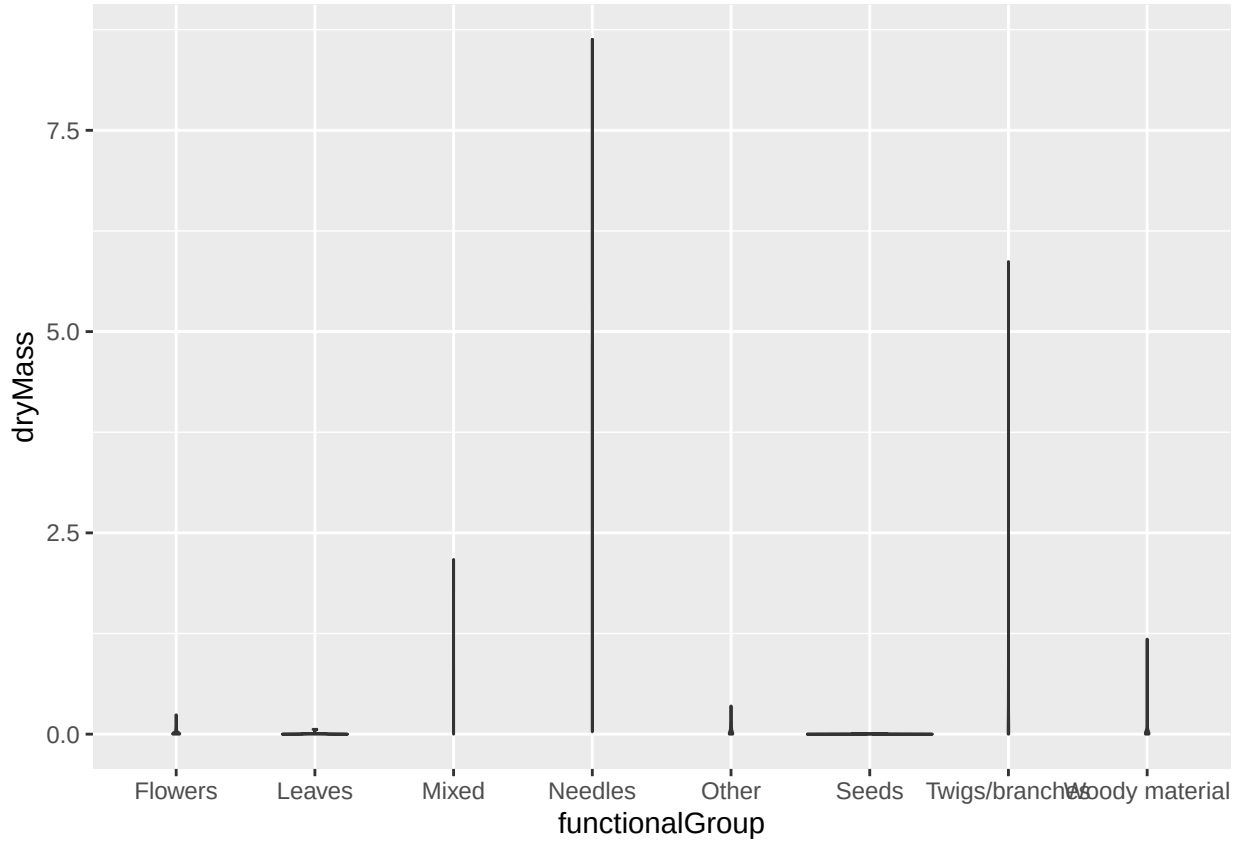
15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(Litter) +
  geom_boxplot((aes(x = functionalGroup, y = dryMass)))
```




```
#this creates the boxplot of dryMass by functional group

ggplot(Litter) +
  geom_violin((aes(x = functionalGroup, y = dryMass)))
```




```
#this creates the violin plot of dryMass by functional group
```

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

> Answer: The boxplot is more effective in this case because we do not need to see the full distribution of the data, the summary statistics (given by the box plot) is enough in this instance.

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

> Answer: Needles tend to have the highest biomass at these sites.