Assignment 3: Data Exploration

Fiona Kelley

Fall 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.

```
#Installing and loading necessary packages.
#install.packages("dplyr")
#install.packages("tidyverse")
#install.packages("lubridate")
#install.packages("ggplot2")
library(dplyr)
library(tidyverse)
library(lubridate)
library(ggplot2)
```

```r
#Imports the required datasets and assigns new names.
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)
```

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: According to the National Library of Medicine, ecotoxicology of neonicotinoids is of great interest because the insecticides impact non-target species, including vital pollinators and invertebrates. The toxicity of the insecticides can spread beyond agricultural realms into natural habitats, altering essential ecological processes.

   Source: Effects of neonicotinoids and fipronil on non-target invertebrates (ncbi.nlm.nih.gov)

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: Studying litter and woody debris that falls to the ground in forests may assist in evaluating forest productivity through decomposition processes. Such processes provide vital ecosystem services in carbon sequestration and nutrient cycling.

   Source: Woody Debris (fs.usda.gov)

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. Litter and woody debris sampling was carried out at terrestrial NEON sites, with woody vegetation exceeding 2 meters in height. 2. Sampling occurred only in tower plot locations that were selected randomly within the 90% flux tower footprint of primary and secondary airsheds.
   3. Ground traps were sampled once annually. Frequent sampling occurred in deciduous forest sites and less frequent sampling was carried out at evergreen sites.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
#The dim() function determines how many rows and columns are in a dataset.
dim(Neonics)
```

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

```r
#Neonics has 4623 rows and 30 columns.
```

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?

```r
#This summarizes the data in the "Effect" column for the Neonics dataset.
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
```

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

Answer: The most common effects might be of specific interest because these highlight the well-established and studied relationships between insect species population, mortality, and behavior and neonicotinoids. Such effects highlight the direct and noticable impacts of the pesticide.

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...]

```r
#Summarizes Species.Common.Name column.
Summary_CommonSpecies <- summary(Neonics$Species.Common.Name)
#Sorts column from most studied to least studied species.
Sorted_Species <- sort(Summary_CommonSpecies, decreasing = TRUE)
#Determines the six most commonly studied species.
MostCommonSpecies <- head(Sorted_Species)
MostCommonSpecies
```

```
##               (Other)             Honey Bee        Parasitic Wasp
##                   670                   667                   285
## Buff Tailed Bumblebee   Carniolan Honey Bee            Bumble Bee
##                   183                   152                   140
```

Answer: The most commonly studied species in the dataset is entitled "other", followed by the Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, and the Bumble Bee. Bee species are likely of specific interest because of the pollination roles carried out by the insects. Additionally, Parasitic Wasps are critical for targetting pests and invasive species.

Source: Parasoid Wasps (extension.umn.edu)

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?

```
#Determines the class of the column.
class(Neonics$Conc.1..Author.)
```

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

Answer: The column Conc.1..Author is factor, not numeric, because of the "stringAsFactors" command used when importing the dataset. This changed all class types in the dataset to factors.

Explore your data graphically (Neonics)

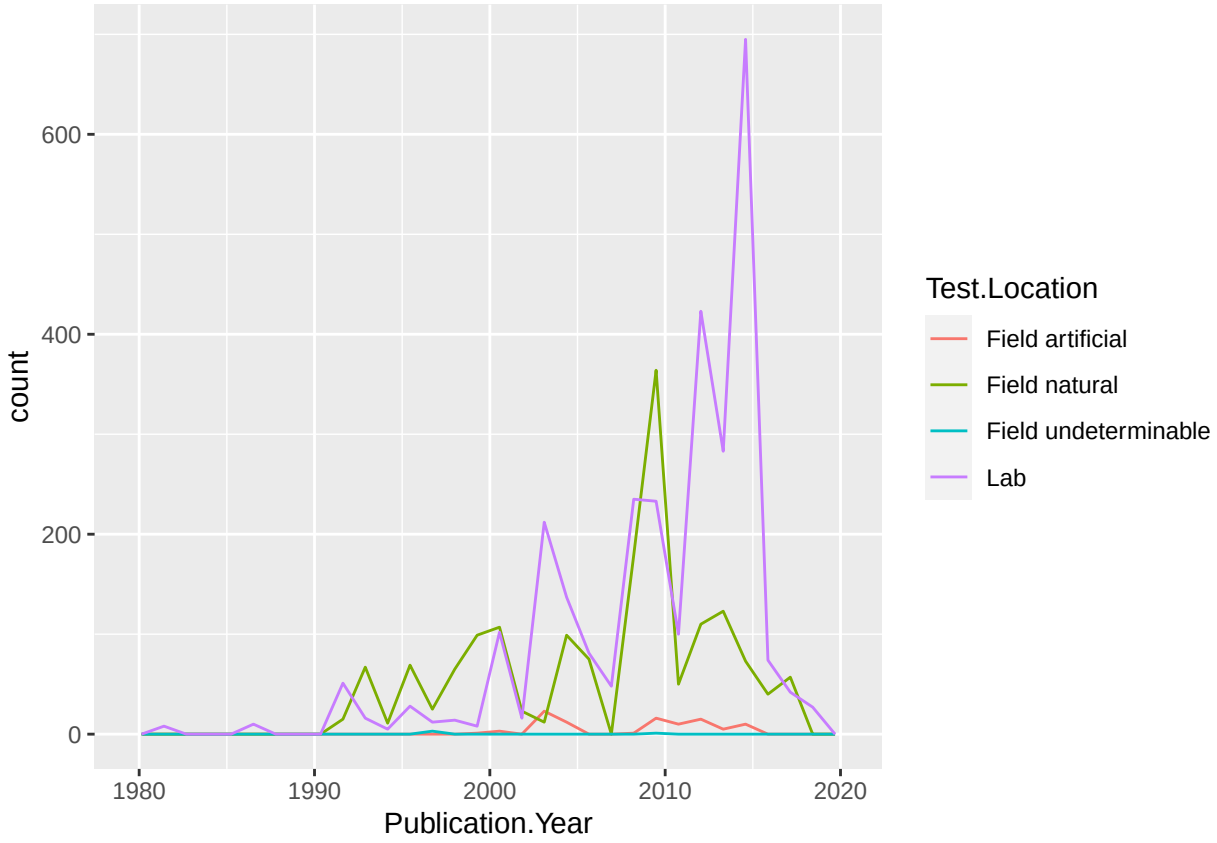
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
#Creates a line graph for Publication Year.
PublicationYear_LineGraph <- ggplot(Neonics,aes(x = Publication.Year)) + geom_freqpoly()
```

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

```
#Colors were added to differentiate publication year data by test location.
ggplot(Neonics, aes(x = Publication.Year, color = Test.Location)) +geom_freqpoly()
```

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



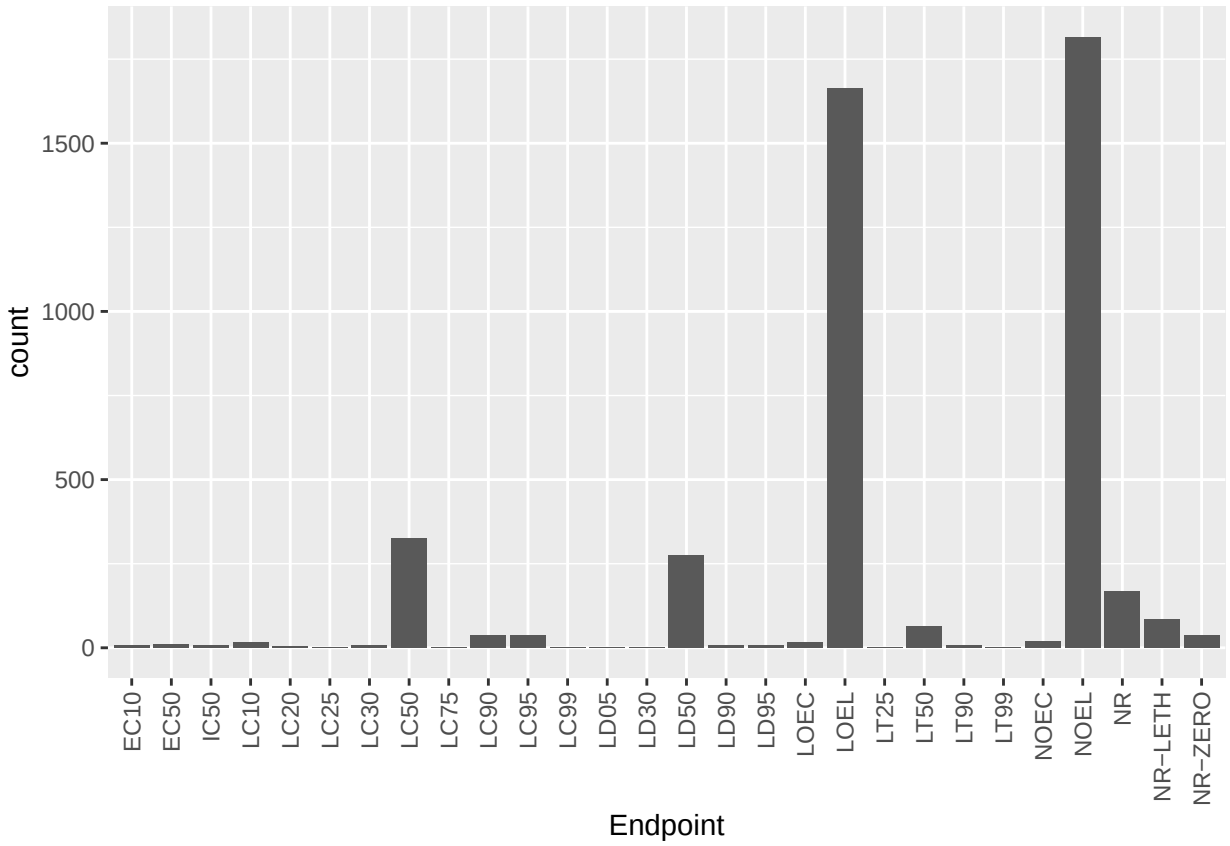
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are the "Lab" and "Field Natural" locations. During the 21st century, publications at the "Lab" location significantly spiked, peaking around 2015. The "Field Natural" location was more prevalent in the late 1900s and early 2000s, with a sudden peak right before 2010.

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...]

```
#Generates a bar graph for the Enpoint counts.
ggplot(Neonics, aes(x = Endpoint)) + geom_bar() +
  theme(axis.text.x = element_text(angle = 90,vjust = 0.5, hjust=1))
```



> Answer: The two most common endpoints are NOEL (No-observable-effect-level) and LOEL (Lowest-observable-effect-level). NOEL indicates there was no significant responses with the highest dose. LOEL identifies significantly different results produced from the lowest dose.

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.

```
#This first class check displays the dataset as a factor.
class(Litter$collectDate)
```

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

```
#The as.Date function converts the dataset to a date class.
Litter$collectDate <- as.Date(Litter$collectDate)
#Checking the class again to confirm the as.Date function worked and the dataset is now a date.
class(Litter$collectDate)
```

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

```
#The unique function determines the two sampled dates in August 2018.
August2018_Dates <- unique(Litter$collectDate[month(Litter$collectDate) == 8])
August2018_Dates
```

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

```
#Summarizes the namedLocation column for the Litter dataset.
summary(Litter$namedLocation)
```

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

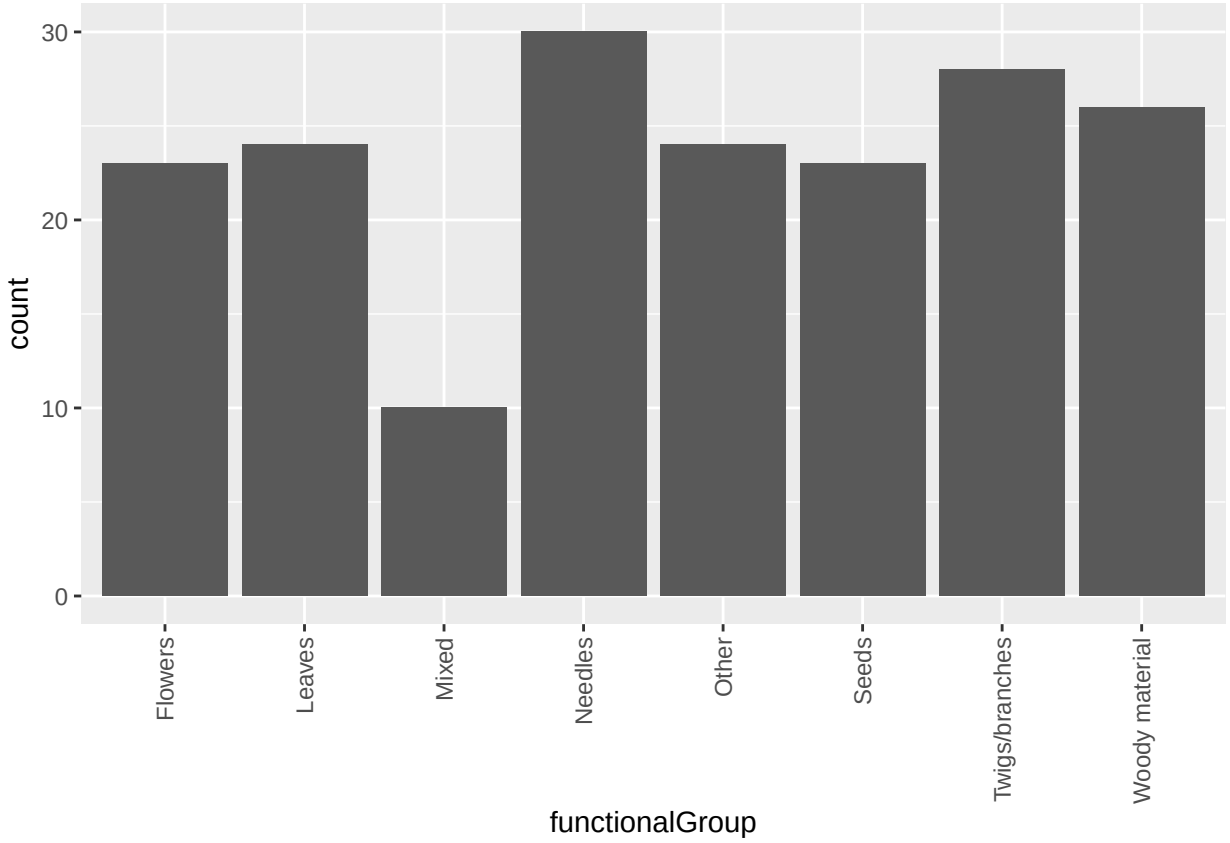
```
#This determines the number of plots sampled at Niwot Ridge (12 plots).
NiwotRidge_PlotNumber <- unique(Litter$namedLocation)
NiwotRidge_PlotNumber
```

```
##  [1] NIWO_061.basePlot.ltr NIWO_064.basePlot.ltr NIWO_067.basePlot.ltr
##  [4] NIWO_040.basePlot.ltr NIWO_041.basePlot.ltr NIWO_063.basePlot.ltr
##  [7] NIWO_047.basePlot.ltr NIWO_051.basePlot.ltr NIWO_058.basePlot.ltr
## [10] NIWO_046.basePlot.ltr NIWO_062.basePlot.ltr NIWO_057.basePlot.ltr
## 12 Levels: NIWO_040.basePlot.ltr ... NIWO_067.basePlot.ltr
```

Answer: The summary function details the count of each plot in the dataset; for example the NIWO_040 plot had a count of 20 observations in the namedLocation column. The unique function produces the number of individual plots within the column. Therefore, the function identified 12 unique plots sampled at Niwot Ridge.

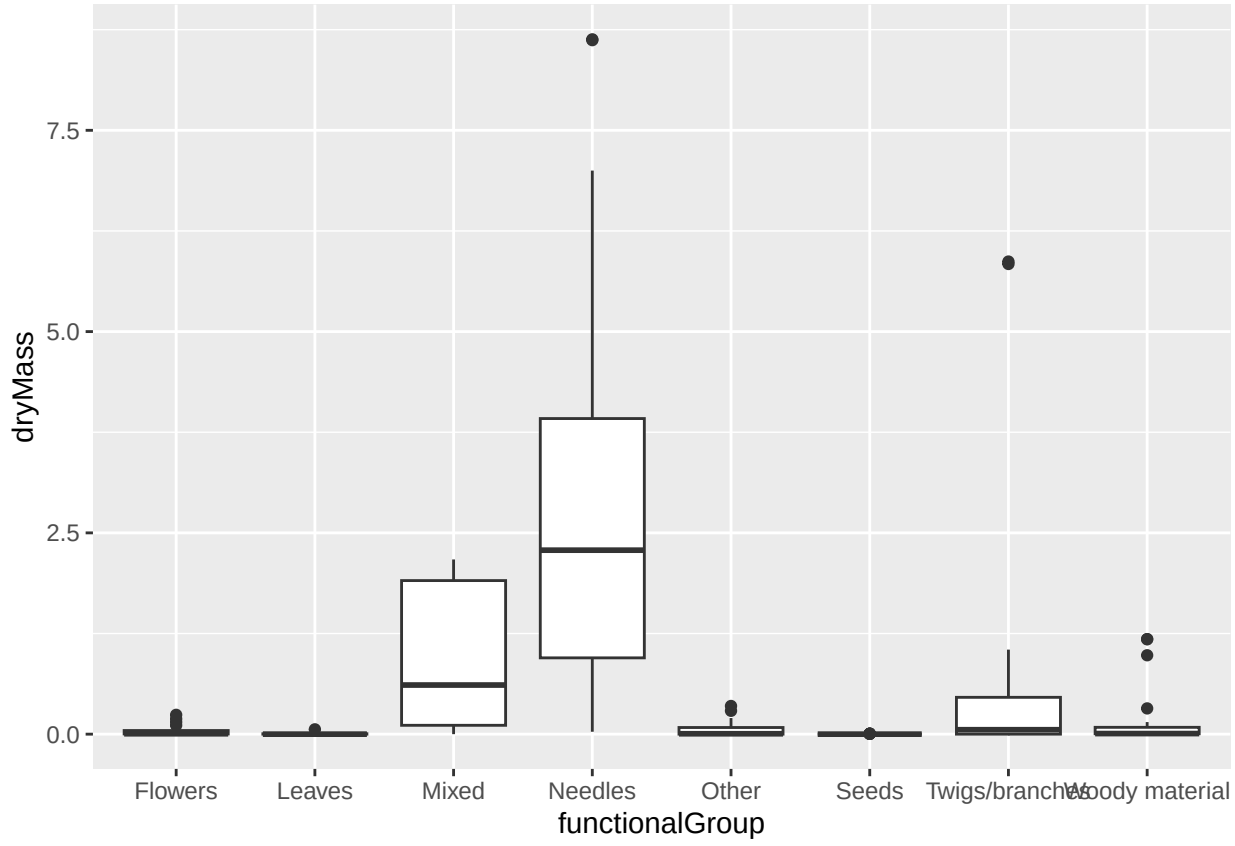
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#This produces a bar graph for the functionalGroup column in the Litter dataset.
ggplot(Litter, aes(x = functionalGroup)) + geom_bar() + theme(axis.text.x = element_text(angle = 90,vjus
```

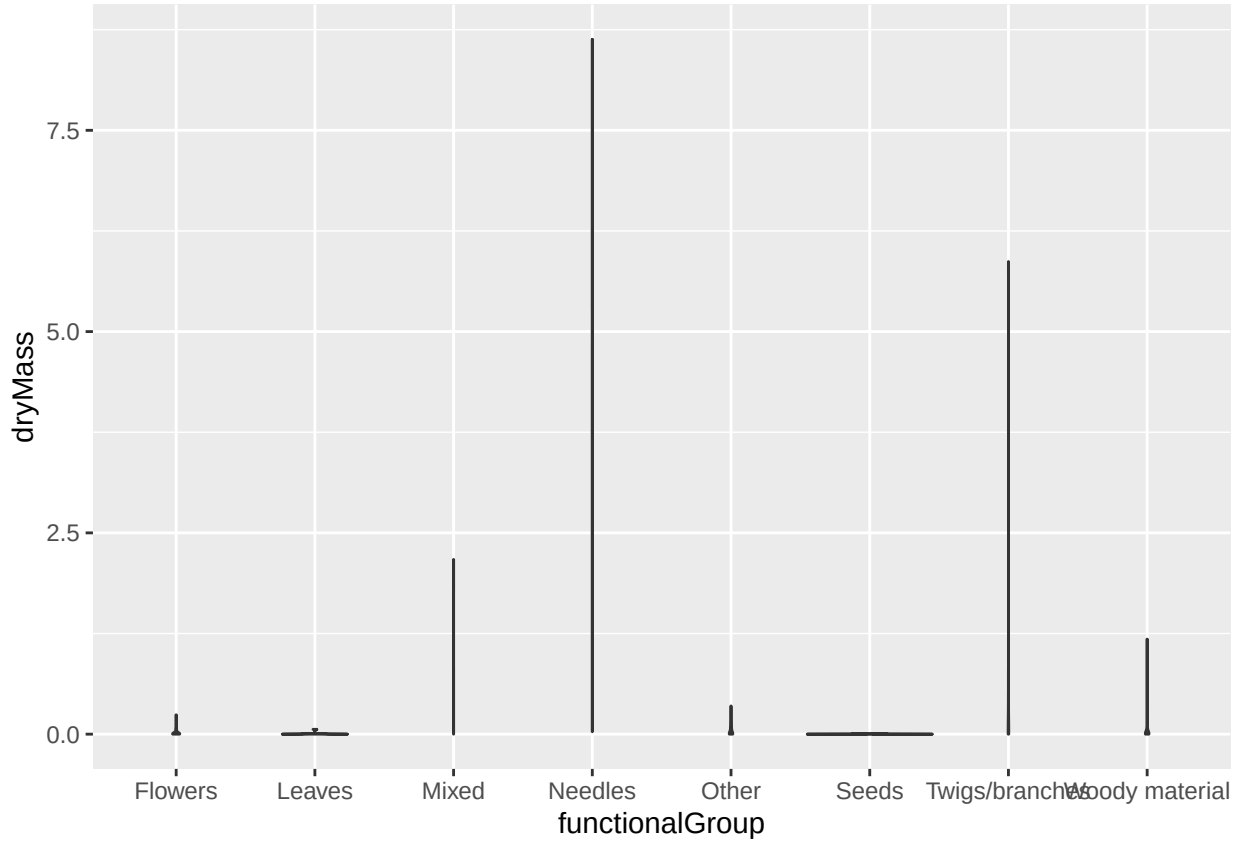


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#This produces a boxplot for dryMass by functionalGroup.
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) + geom_boxplot()
```




```
#The produces a violin plot for dryMass by functionalGroup.
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) + geom_violin()
```



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

> Answer: The boxplot is a more effective visualization option than the violin plot because the boxplot easily conveys summary statistics and outliers. A violin plot for the data distribution in this case is unsuitable because the small sample size and outliers are affecting the range and disorting the visualization.

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

> Answer: At these sites, "Needles" and "Mixed" have the highest biomass. As seen in the boxplot, these two litter types have higher median and interquartile values, indicating higher biomass values collected.