Assignment 3: Data Exploration

Kim Myers

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 Tuesday, January 28 at 1:00 pm.

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.

```
getwd()
```

```
## [1] "C:/Users/Temp/Documents/Duke/S20/DataAnalytics/Environmental_Data_Analytics_2020/Assignments"
```

```
library(tidyverse)

Neonics <- read.csv("../Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv")
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: While neonicitinoids have a low health risk in humans, they are highly toxic to certain insects. In particular, some studies have found that the insecticide causes honey-bee colony collapse disorder and can affect bird populations by decreasing the amount of insect prey. Results

of these studies are controversial and the scientific community has not yet reached a consensus on whether neonicitinoids are an ecological health risk.

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 woody debris can serve many purposes in forests. First of all, they are a source of nutrients in soil and are a crucial part of the carbon cycle. Additionally, litter and debris act as food sources for decomposers, which are a foundational part of the food chain. There are many more services that forest litter and woody debris provide.

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: * litter and debris are collected in ground and elevated traps * in each 400x400m plot, 4 litter pair (ground and elevated) traps are set * number of plots conducted and placement of traps within the plot depends on land cover type

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

6. Using the `summary` function, 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
```

   Answer: The effects listed in the dataset generally pertain to insect behavior and survival, two elements scientists might be interested in to understand the effect neonicitinoids with have on other organisms.

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.

```
summary <- summary(Neonics$Species.Common.Name)
summhigh <- sort(summary, decreasing = T)
summhigh[1:6]
```

```
##               (Other)             Honey Bee        Parasitic Wasp
```

```
##                  670                  667                  285
## Buff Tailed Bumblebee   Carniolan Honey Bee           Bumble Bee
##                  183                  152                  140
```

Answer: Four of the six most common species in the dataset are bees. These species are important pollinators for foods that humans consume and other species' food sources. The parasitic wasp also acts as pest control in agricultural areas.

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

```
class(Neonics$Conc.1..Author)
```

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

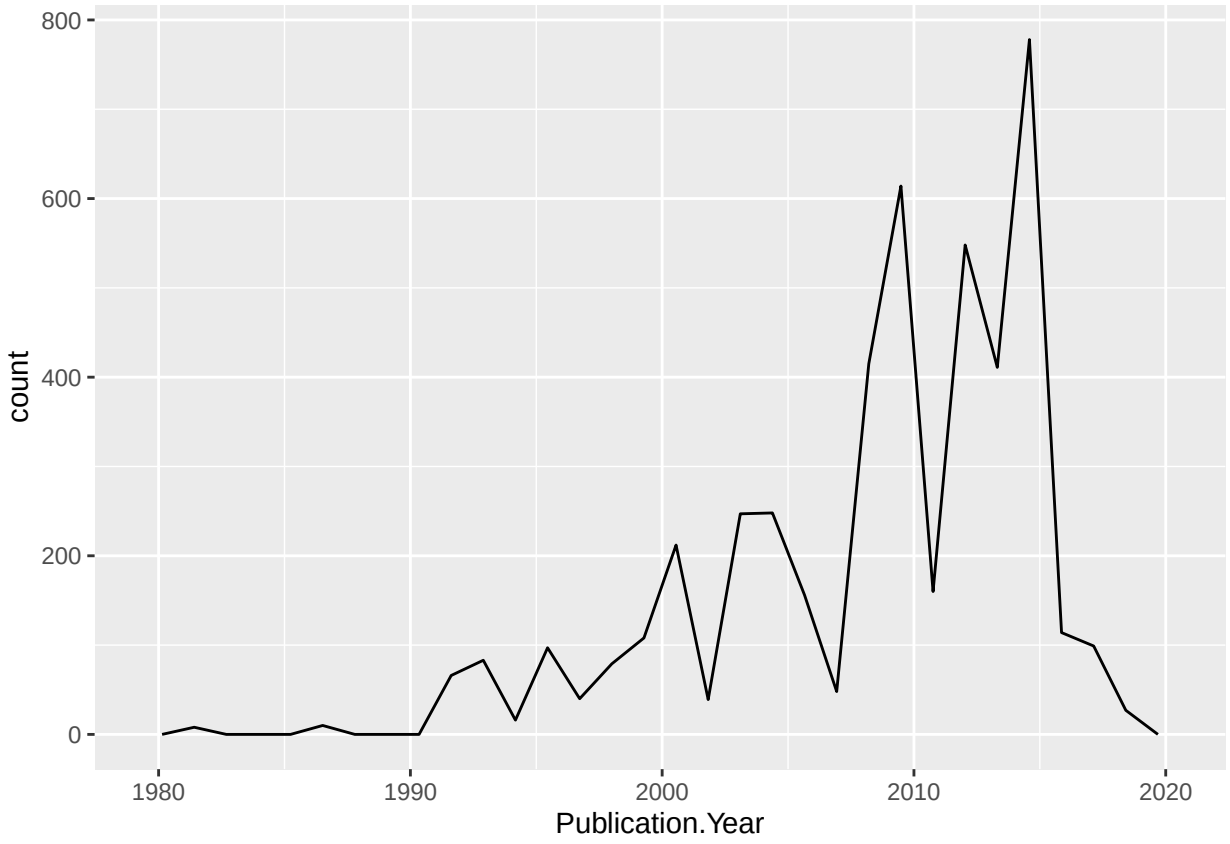
Answer: This field is the concentration of insecticide found in the insect. It's not numeric because some of the cells include text values.

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))
```

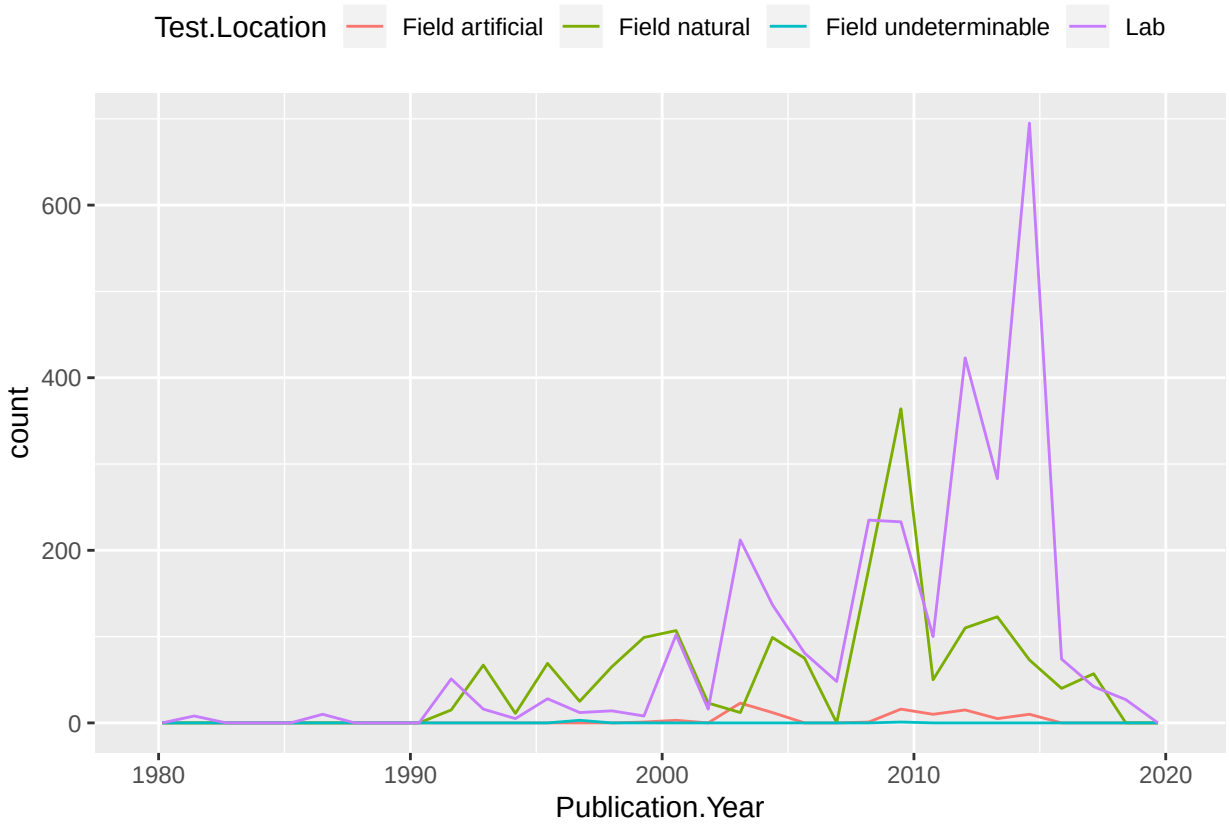
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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)) +
  theme(legend.position = "top")
```

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

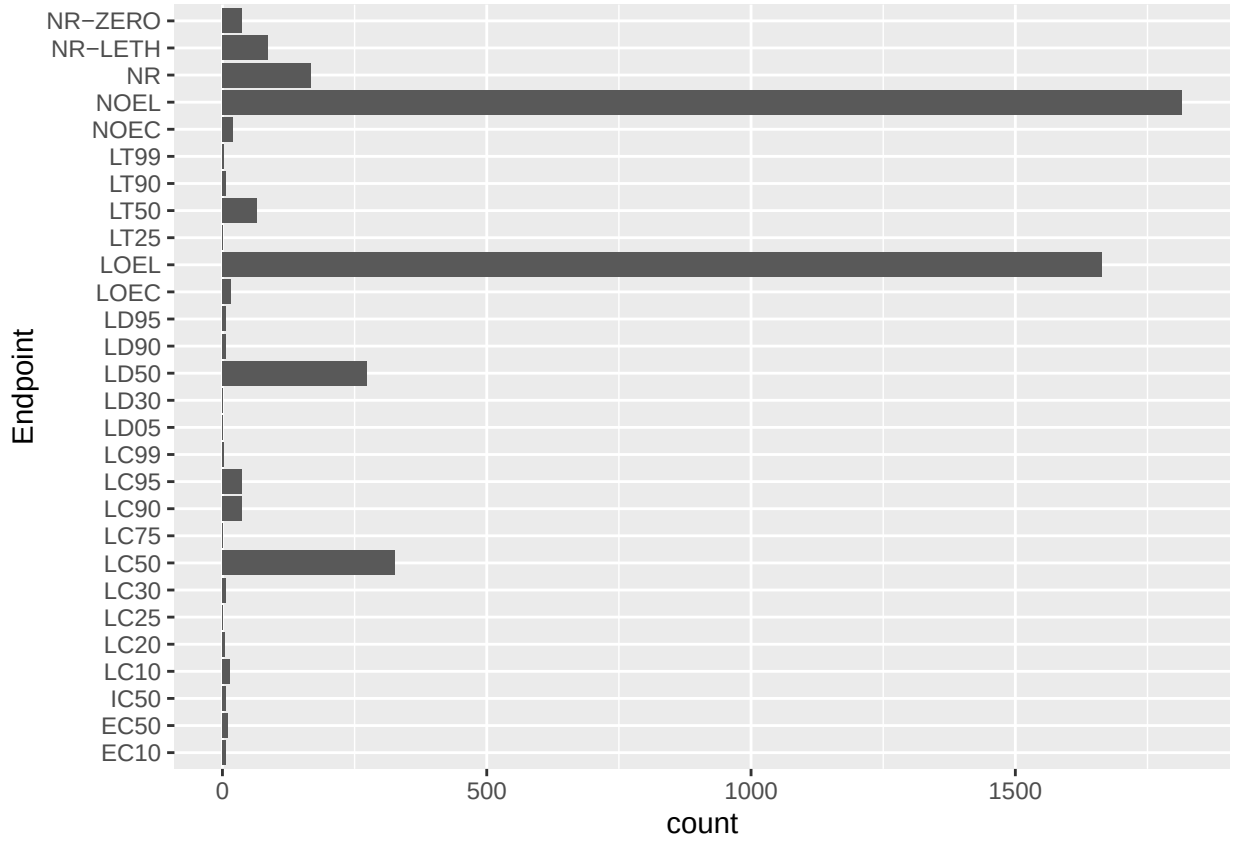


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: Most of the test locations sites were either natural in the field or in the lab. In the early 2000s and throughout the 2010s, lab sites were the most common. From 1990 to 2000 and the late 2000s, natural field sites were the most common.

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.

```
ggplot(Neonics) + geom_bar(aes(x = Endpoint)) +
  coord_flip()
```



Answer: NOEL and LOEL are the two most common endpoints. LOEL means that the lowest dose that produced effects were significantly different from responses to the controls. NOEL refers to a non-observable-effect-level, meaning the effects were not significantly different.

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(Litter$collectDate)
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
class(Litter$collectDate)
```

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

```
unique(Litter$collectDate)
```

```
## [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`?

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

```
##  [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
```

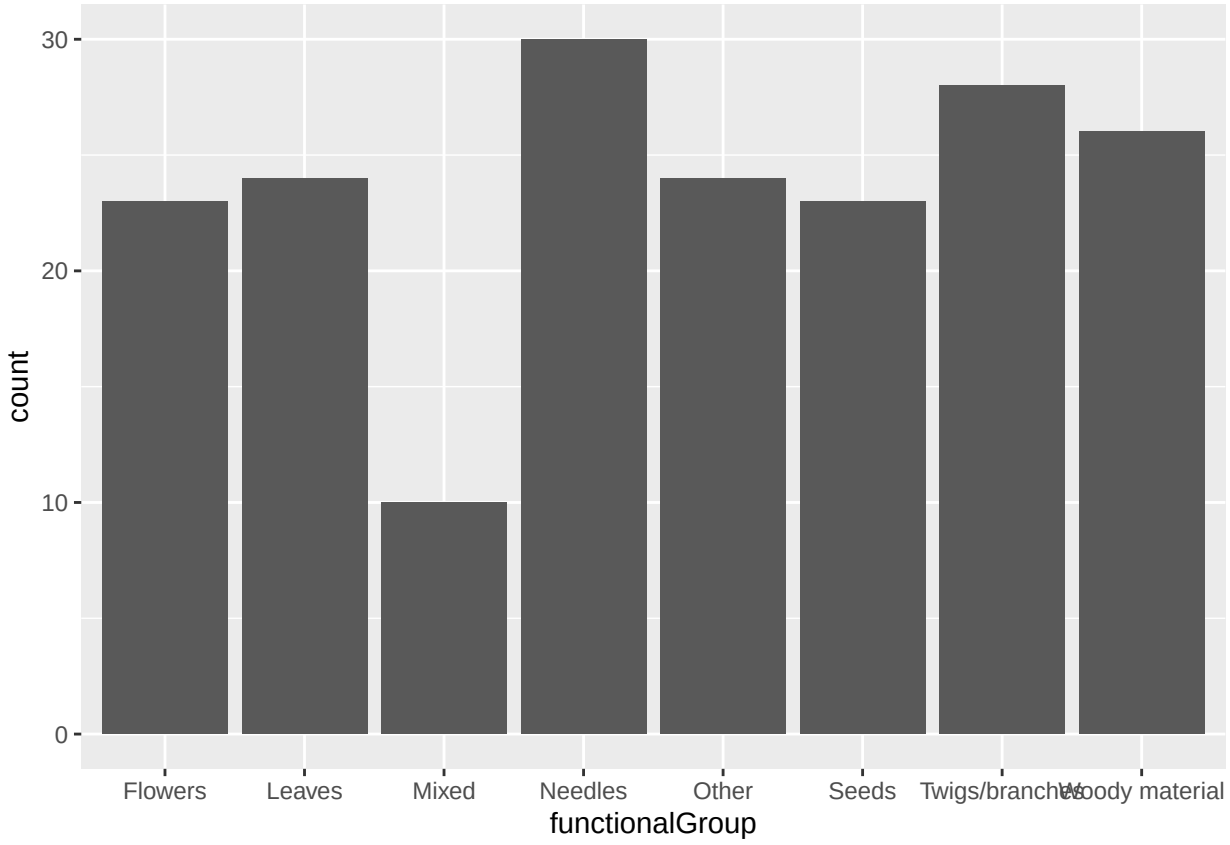
```
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

```
#summary(Litter$plotID)
```

Answer: The unique function lists all unique factors. The summary function also does this but includes the number of rows attributed to each factor.

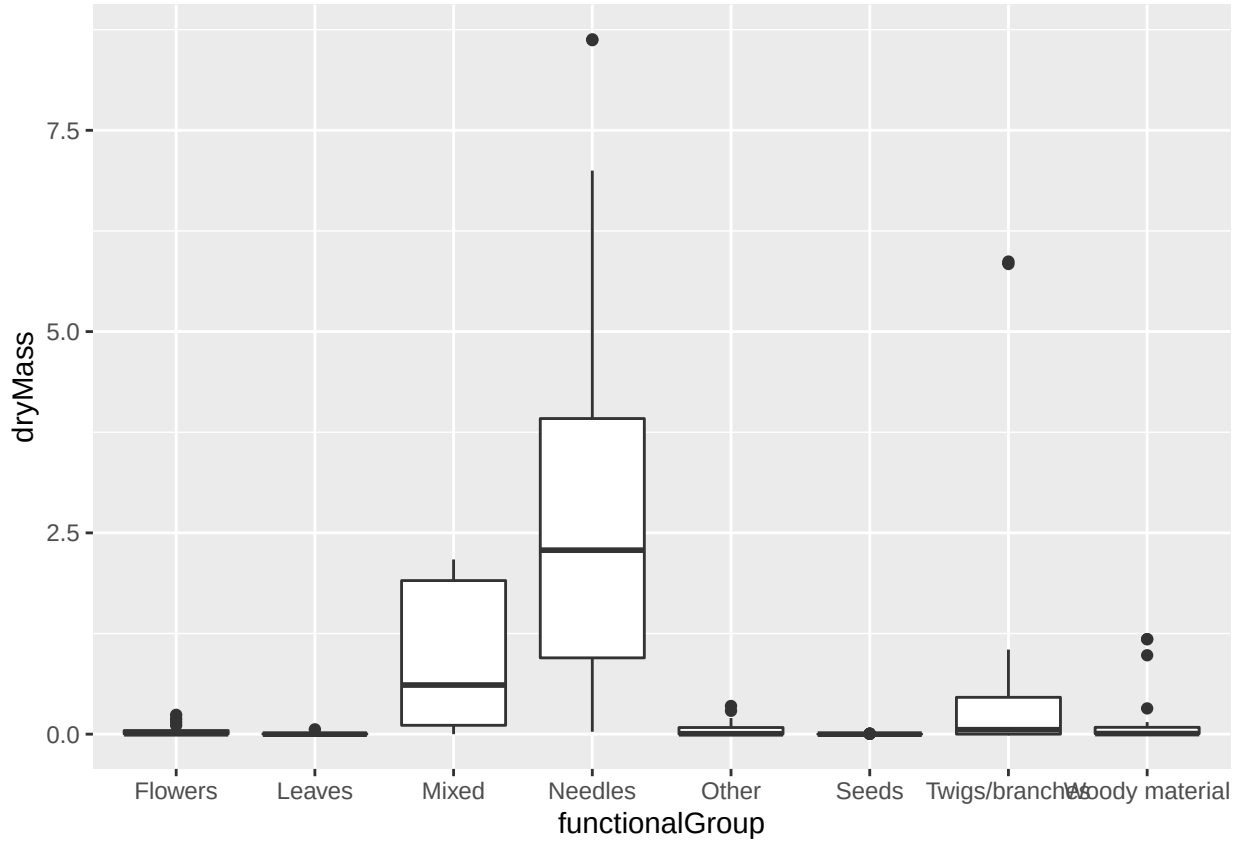
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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



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))
```

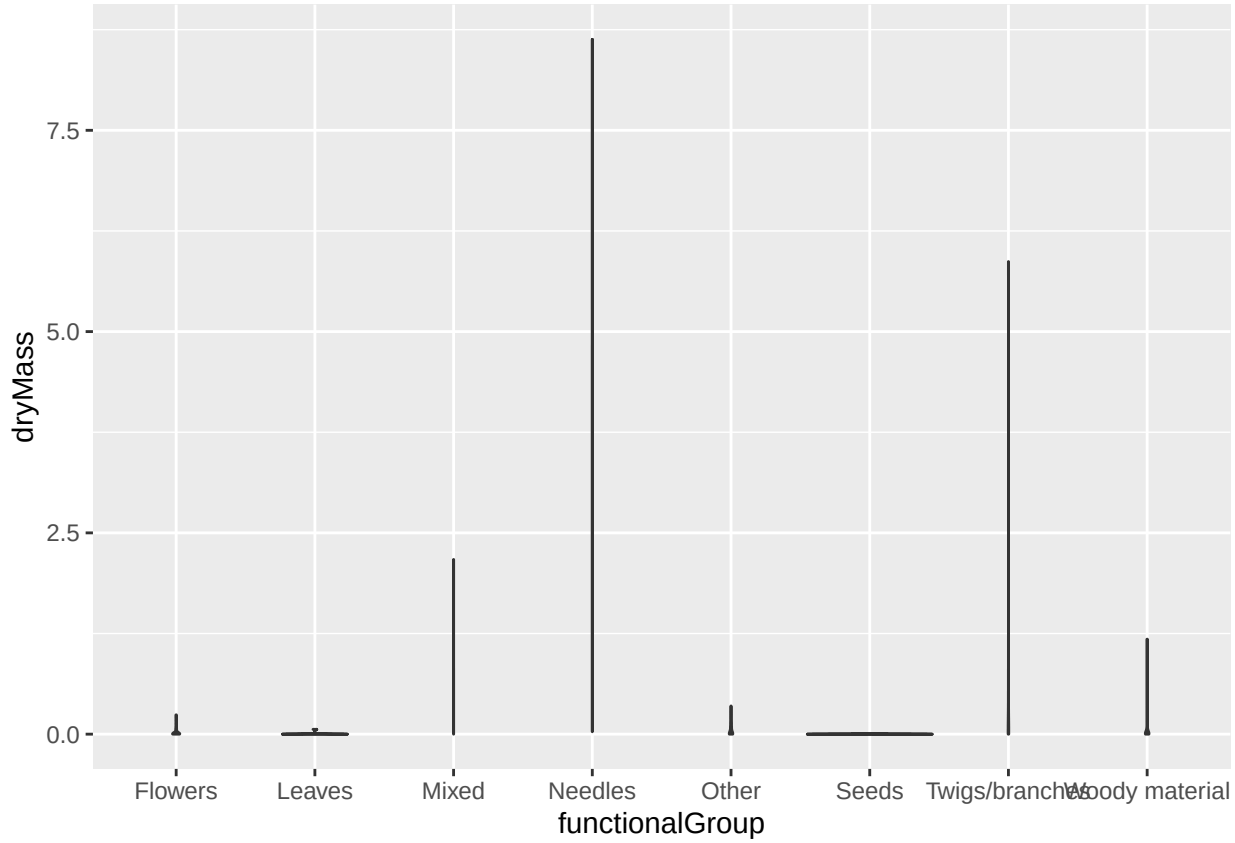



```
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass), draw_quantiles = c(0.25, 0.5, 0.75))
```

```
## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to unique
## 'x' values

## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to unique
## 'x' values

## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to unique
## 'x' values
```



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

> Answer: The violine plot is less effective than the boxplot because the values of each litter type are either very clumped in a small range or values or equally spread across a large range. The boxplot provides a visual representation of value distribution without being horizontally shapes by the data.

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

> Answer: Needles tend to have the highest biomass and mixed litter has the second highest.