Assignment 3: Data Exploration

Christina Li

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.

```
getwd()
```

```
## [1] "C:/Users/li_ch/Desktop/DKU/Year 2/Term 2/Environmental Data Analytics/GIT Hub/Environmental_Data
```

```
library("tidyverse")
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: Despite the harmful insects, beneficial insects such as pollinators are also killed when using neonicotinoids. Therefore, studying the ecotoxicology of this insecticides can help understand its chemical properties and toxicity to insect population, and establish user guide/policy regulating the use of the substance.

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: Tree litter and woody debris are part of the forest carbon storage that provides energy, nutrients and habitat for different organisms during decomposition. Studying tree litter and woody debris can help us understand the carbon flux between the ecosystem and the atmosphere, as well as carbon stock characteristics.

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: *Randomly chosen sampling sites contain woody vegetations >2 meters and only occure in tower plots.* Trap placement within plots may be either targeted or randomized, depending on the vegetation. *Sampling frequency depends on forest type and time of the year: 1) sampling once every 2 weeks in deciduous forest sites during senescence; 2) sampling once every 1-2 months at evergreen sites; 3)groud traps are sampled onece per year; 4) no sampling during dormant season (winter) which last up to 6 months.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) #Row,Col
```

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

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?

```
sort(summary(Neonics$Effect), decreasing = TRUE)[1]
```

```
## Population
##       1803
```

Answer: The most common Effects studied are Population and Mortality. Studying the ecotoxicology of a chemical is evaluating its effect on a populatiton; and for insecticides, we would want to know whether it

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(Neonics$Species.Common.Name)[1:6]
```

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

Answer: The six most common species are Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. These insects are beneficial, and many are susceptible to insecticides but are irreplaceable essential pollinators for crop plants.

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.)
```

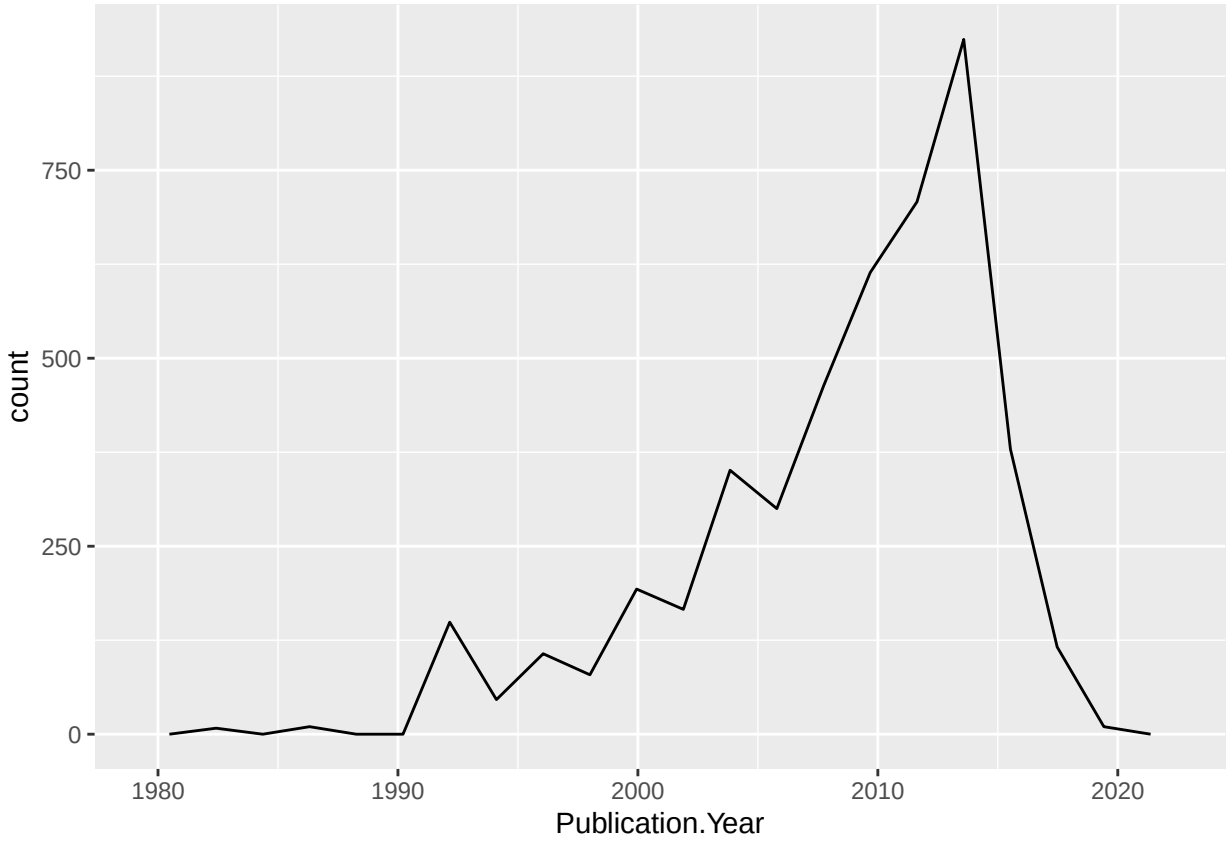
```
## [1] "factor"
```

Answer: Beacuse it is a concentration where the % sign is in the next column. If we treat this column as numeric data, the calculation will be 100 time larger than actural value.

Explore your data graphically (Neonics)

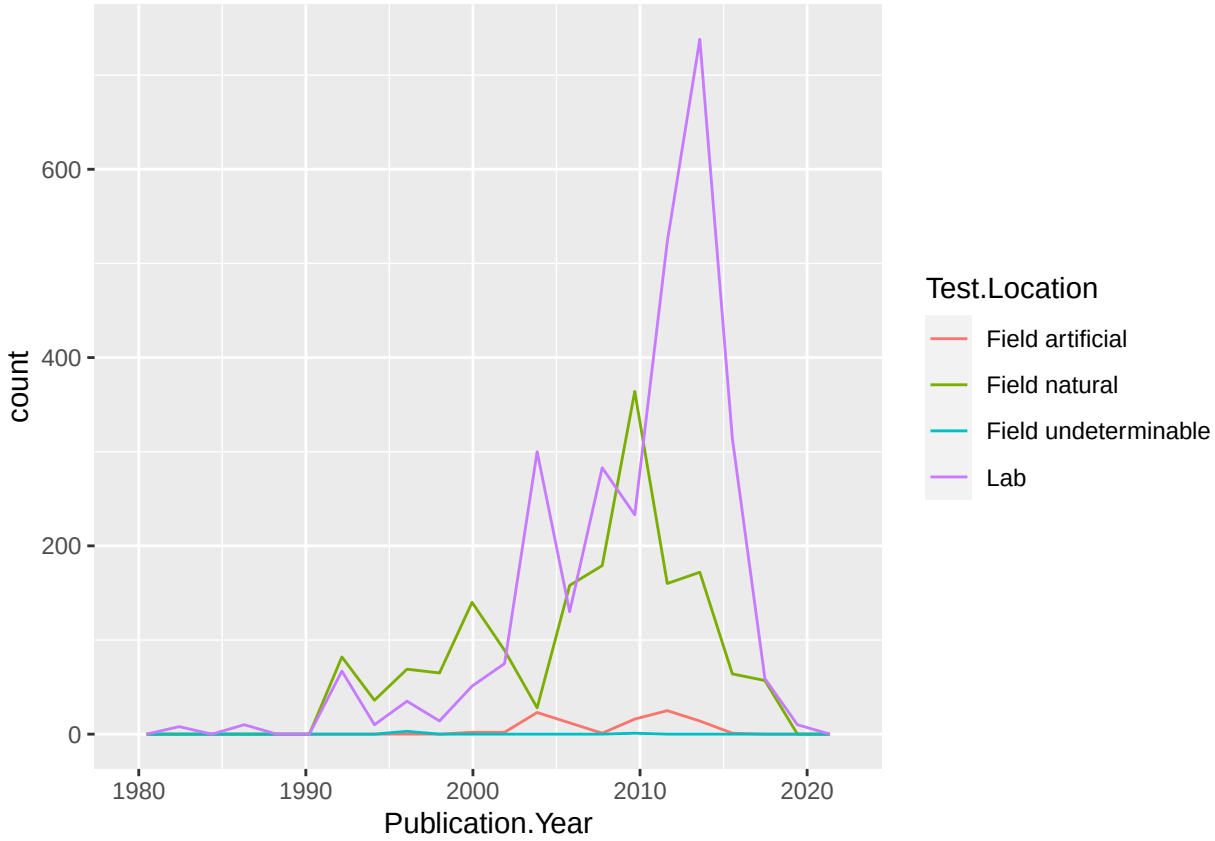
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
library(ggplot2)
ggplot(Neonics)+
  geom_freqpoly(aes(x = Publication.Year), bins = 20)
```



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), bins = 20)
```

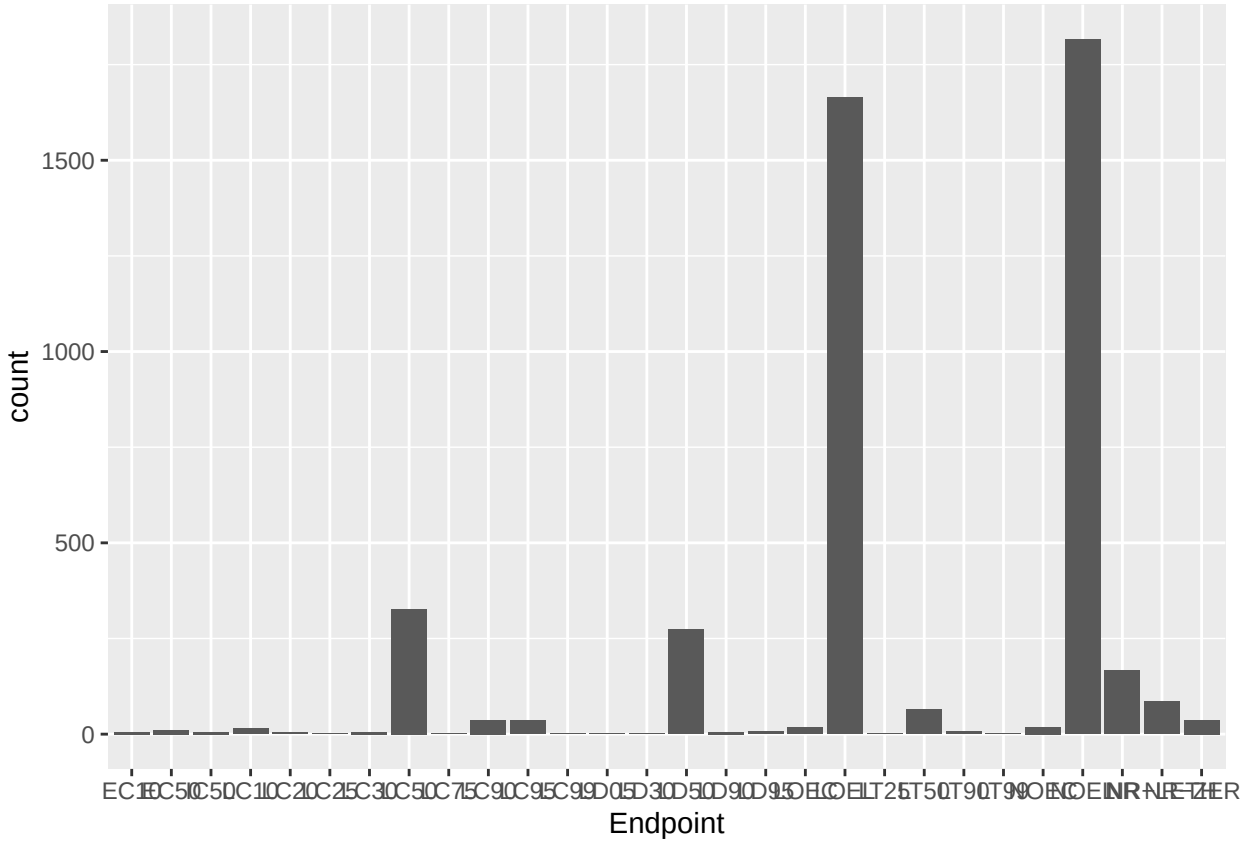


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test location is in the lab. Overall, the lab is the most frequent test location. However, in 2010, the natural field occurs the most as a test location. Artificial fields and undeterminable fields remain in low numbers for the whole sampled times.

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))
```



Answer: The two most common endpoints are LOAEL (Lowest Observed Adverse Effect Level) which is defined as "lowest dose producing an adverse effect"; and NOAEL (No Observed Adverse Effect Level) which is defined as "the highest dose producing no adverse effect." These two factors are used to determine the safe dose of chemical use, especially when we want to avoid the health risk of a certain species of interest (in this case it is the bees).

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)
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
```

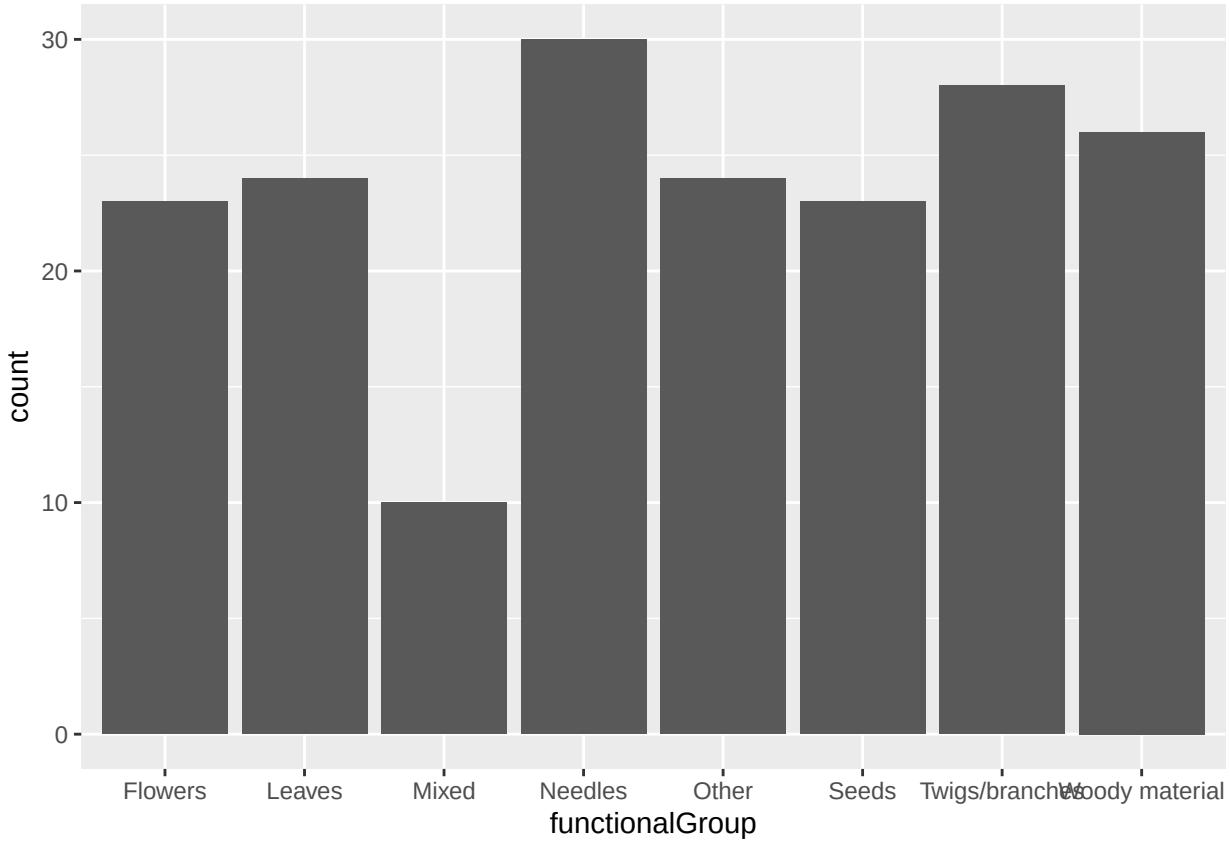
```
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

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

Answer: There are 12 plots sampled. Unique function only gives the name of the unique site and count the number of unique sites; while summary funtion give the name of the sites and how many times each site appears.

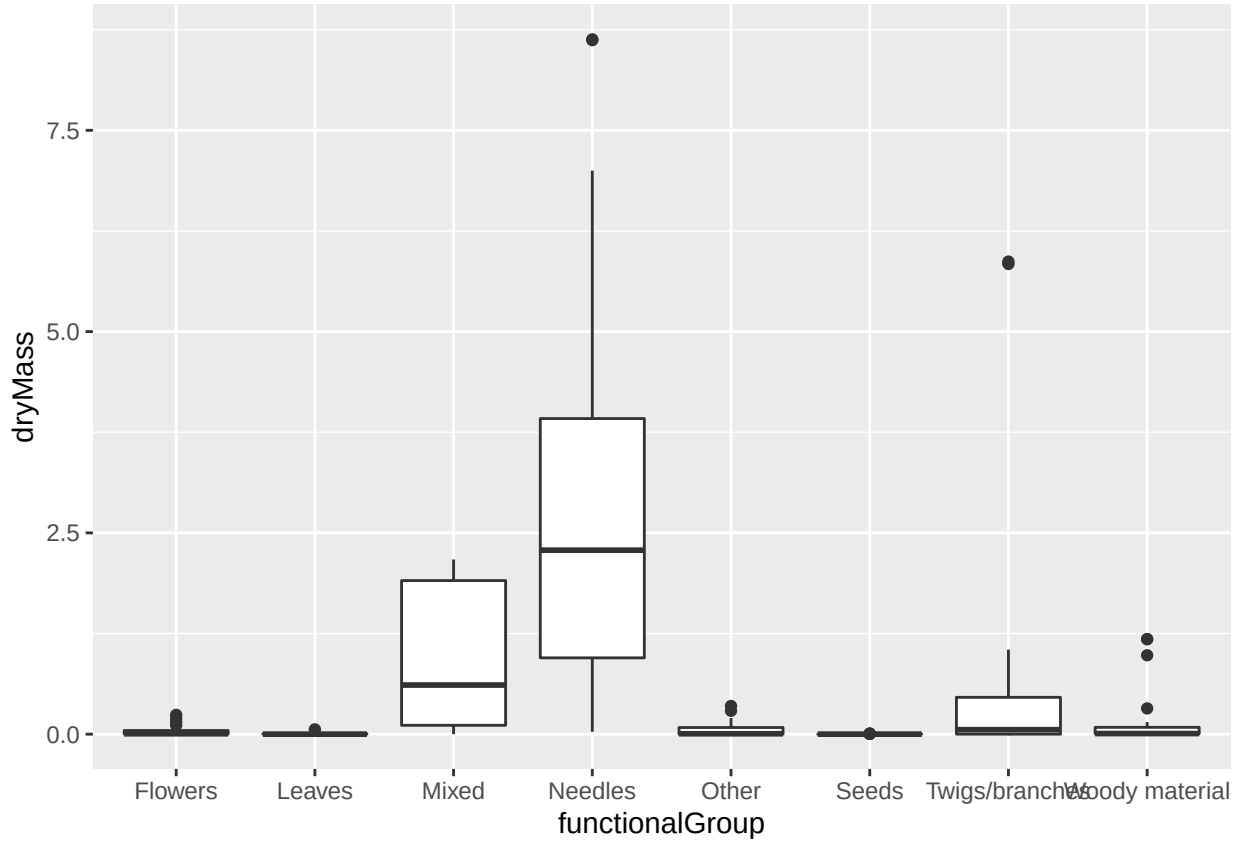
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

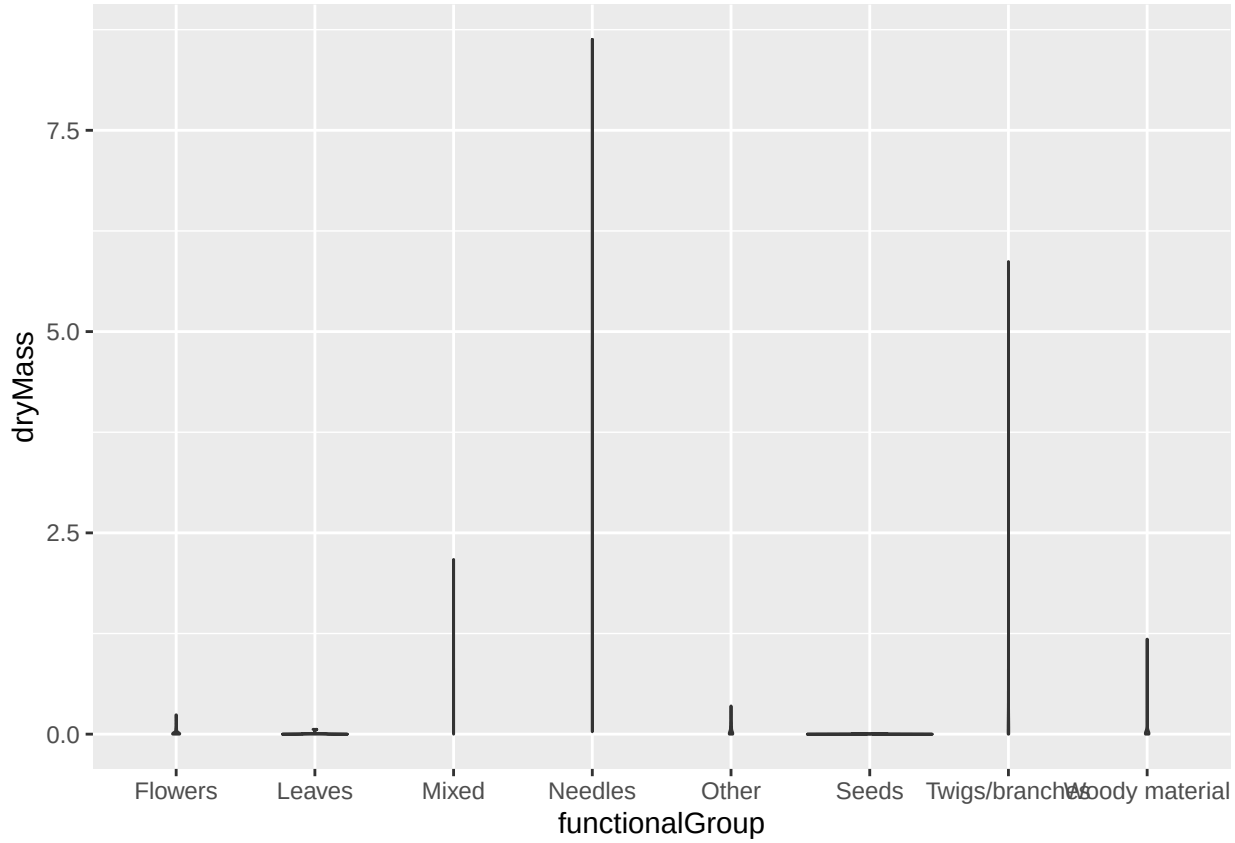


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




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



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

Answer: the boxplot display the skewness in this case but the violin plot fails to do so.

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

Answer: Needles