Assignment 3: Data Exploration

Zhiyuan Chen

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. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.
6. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai.

The completed exercise is due on Sept 30th.

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. Be sure to include the subcommand to read strings in as factors.

```
getwd()
```

```
## [1] "D:/Rfiles/EDA-Fall2022"
```

```
# install.packages('tidyverse')
library(tidyverse)
Neonics.Data <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = TRUE)
Litter.Data <- 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: Neonicotinoids are used for pest control. Initially neonicotinoids were praised for their low-toxicity to many beneficial insects, including bees.However, new research points to potential toxicity to bees and other beneficial insects through low level contamination of nectar and pollen with neonicotinoid insecticides used in agriculture. To measure the actual impact of neonicotinoid insecticides on bees and other beneficial insects, we should know the ecotoxicology of neonicotinoids on insects.

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: Plant debris, defined as woody debris and litter, plays an important role in the carbon (C) cycling of forest ecosystems. Forest litter is the largest source of organic matter in the organic layer and mineral soils, and woody debris provides energy, nutrient, and habitat for microbes during decomposition. To determine the influence of litter and woody debris quality on decomposition and associated nutrient release, we should study litter and woody debris that falls to the ground in forests.

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 fine woody debris sampling is executed at terrestrial NEON sites that contain woody vegetation >2m tall. 2.A 1m buffer around the edge of the plot and all nested subplots are excluded from consideration for sampling to avoid interfering with plant diversity measurements, vegetation structure measurements and areas subject to high traffic around the edge of the plot. 3. Ground traps are sampled once per year. Target sampling frequency for elevated traps varies by vegetation present at the site, with frequent sampling (1x every 2weeks) in deciduous forest sites during senescence, and in frequent year-round sampling (1x every 1-2 months) at evergreen sites.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics.Data)
```

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

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?

```
all_effects <- summary(Neonics.Data$Effect)
all_effects
```

```
##     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
```

```
most_common_effects <- summary(Neonics.Data$Effect, 2)
most_common_effects
```

```
## Population    (Other)
##       1803       2820
```

Answer: The most common effects that are studied are population(1803). These effects can indicate the impacts of neonicotinoids, such as causing death of bees and other beneficial insects.

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.

```
common_name7_sum <- summary(Neonics.Data$Species.Common.Name, 7)
common_name7_sum
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee      Italian Honeybee
##                   152                   140                   113
##               (Other)
##                  3083
```

Answer: They are "Honey Bee", "Parasitic Wasp", "Buff Tailed Bumblebee", "Carniolan Honey Bee", "Bumble Bee", "Italian Honeybee". They're both members of the insect order Hymenoptera. Because by studying these six species, we can know whether neonicotinoids have done harm to beneficial insects, such as bees.

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.Data$Conc.1..Author.)
```

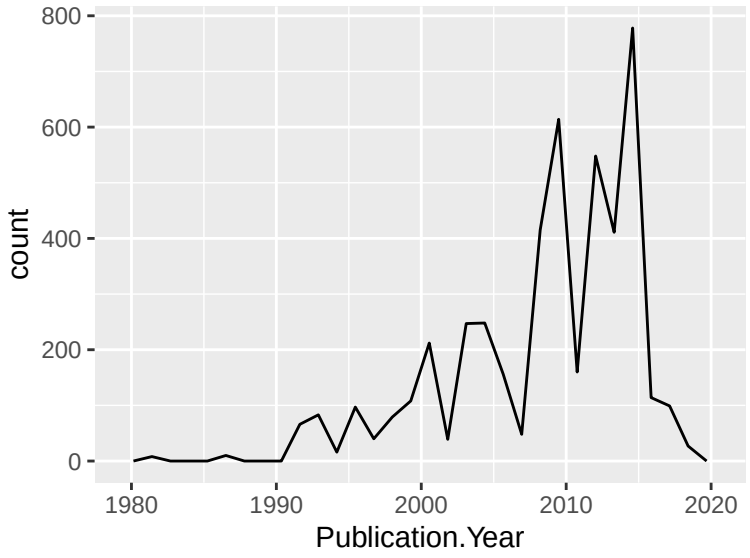
```
## [1] "factor"
```

Answer: The class of Conc.1..Author is factor. Because the data in Conc.1..Author is just some information like zipcodes.

Explore your data graphically (Neonics)

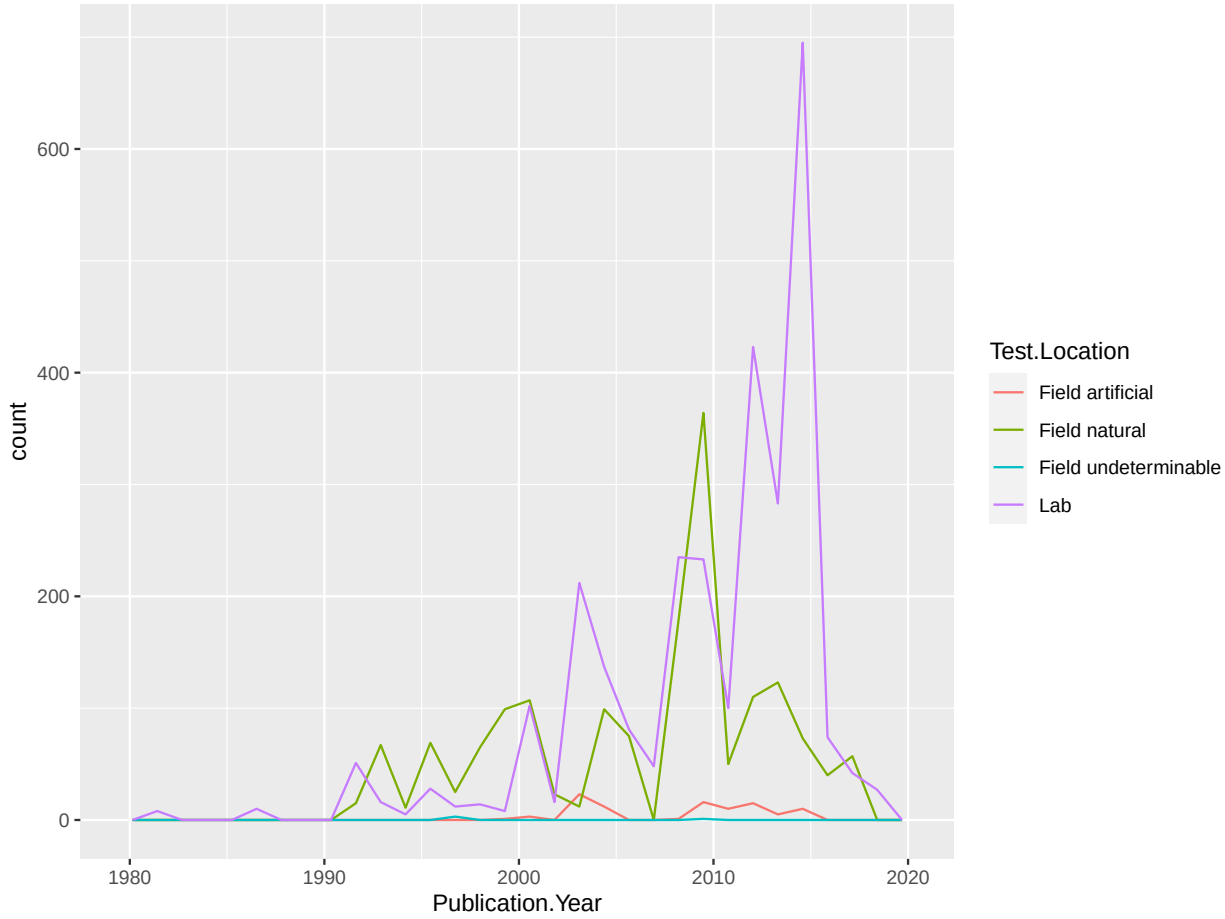
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics.Data) + geom_freqpoly(aes(x = Publication.Year), bins = 30)
```



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

```
ggplot(Neonics.Data) + geom_freqpoly(aes(x = Publication.Year, color = Test.Location),
    bins = 30)
```

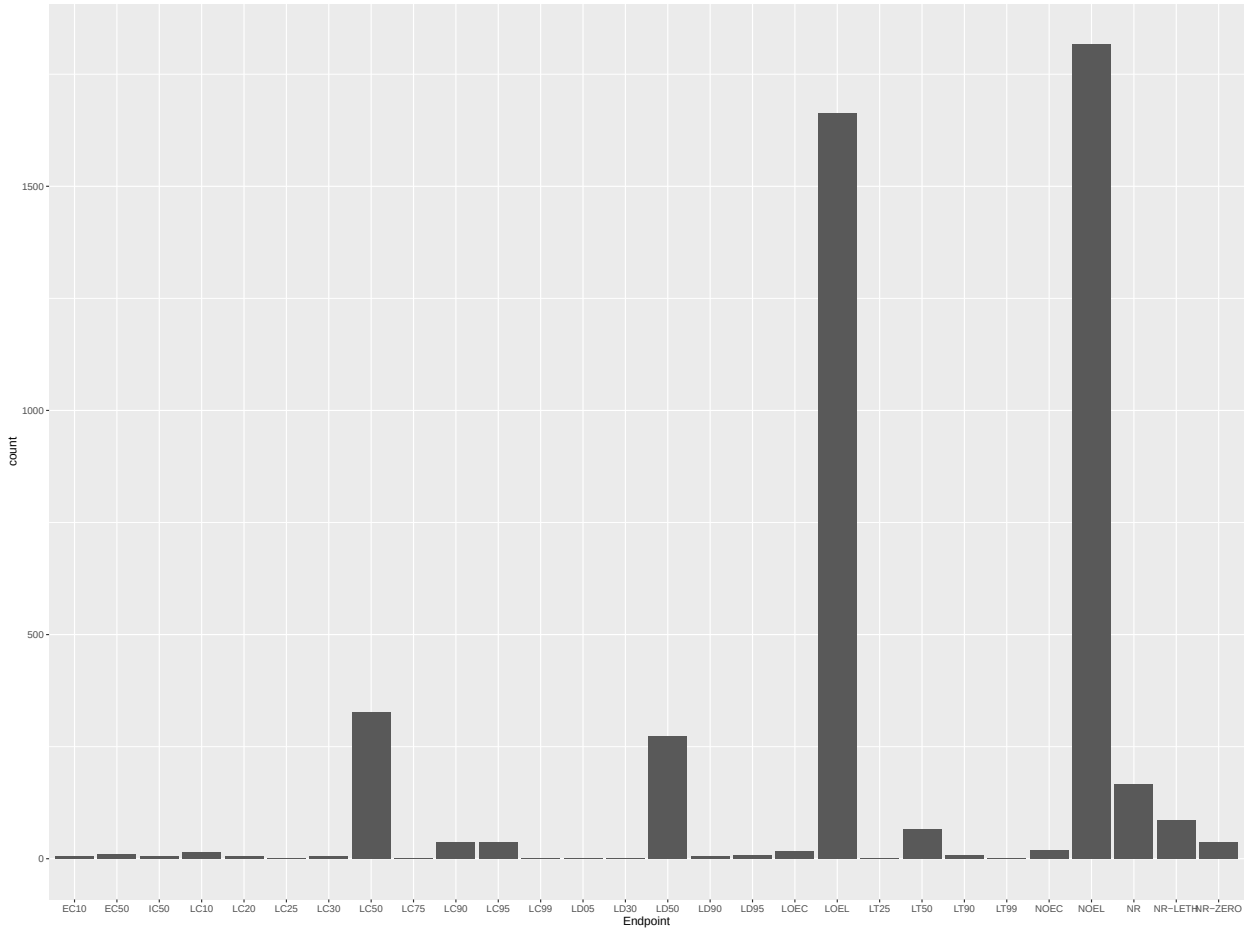


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are labs. They do differ over time. From 1980 to 1990, the lab was the most common test location. During most years of the 1990s, the natural field was the most common test location. While entering the 2000s, the lab took the lead at first. The 2010s witnessed a dramatic increase in lab test frequency.

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.Data) + geom_bar(aes(x = Endpoint))
```



Answer: NOEL and LOEL. No-observable-effect-level(NOEL): highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test. Lowest-observable-effect-level(LOEL): lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls.

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.Data$collectDate)
```

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

```
Litter.Data$collectDate <- as.Date(Litter.Data$collectDate)
class(Litter.Data$collectDate)
```

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

```
unique(Litter.Data$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.Data$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
```

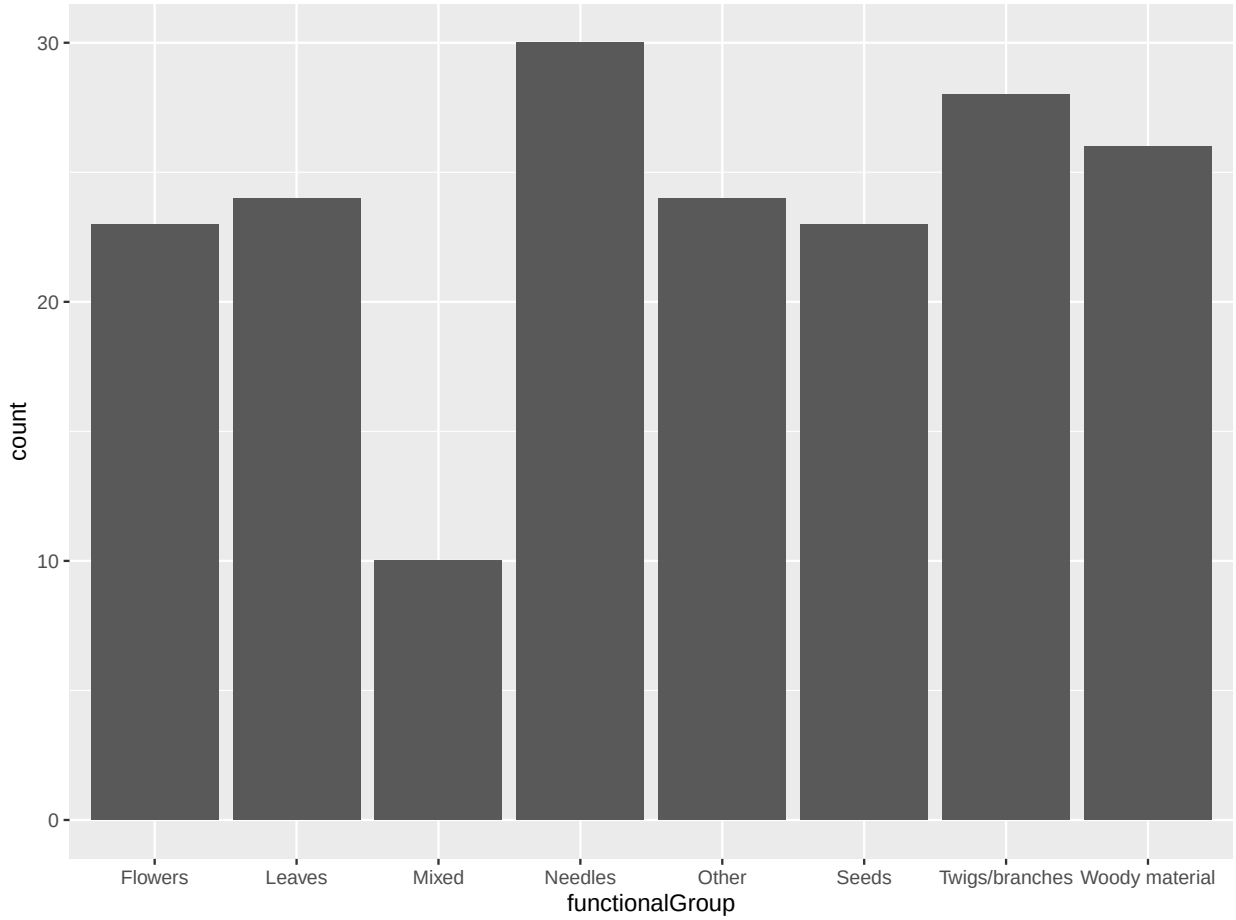
```
summary(Litter.Data$plotID)
```

```
## 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: THe `unique` function just shows 12 levels of plots. But the `summary` function shows how many times a plot were sampled in addition to the 12 levels or IDs, for example NIWO_040 has 20 records.

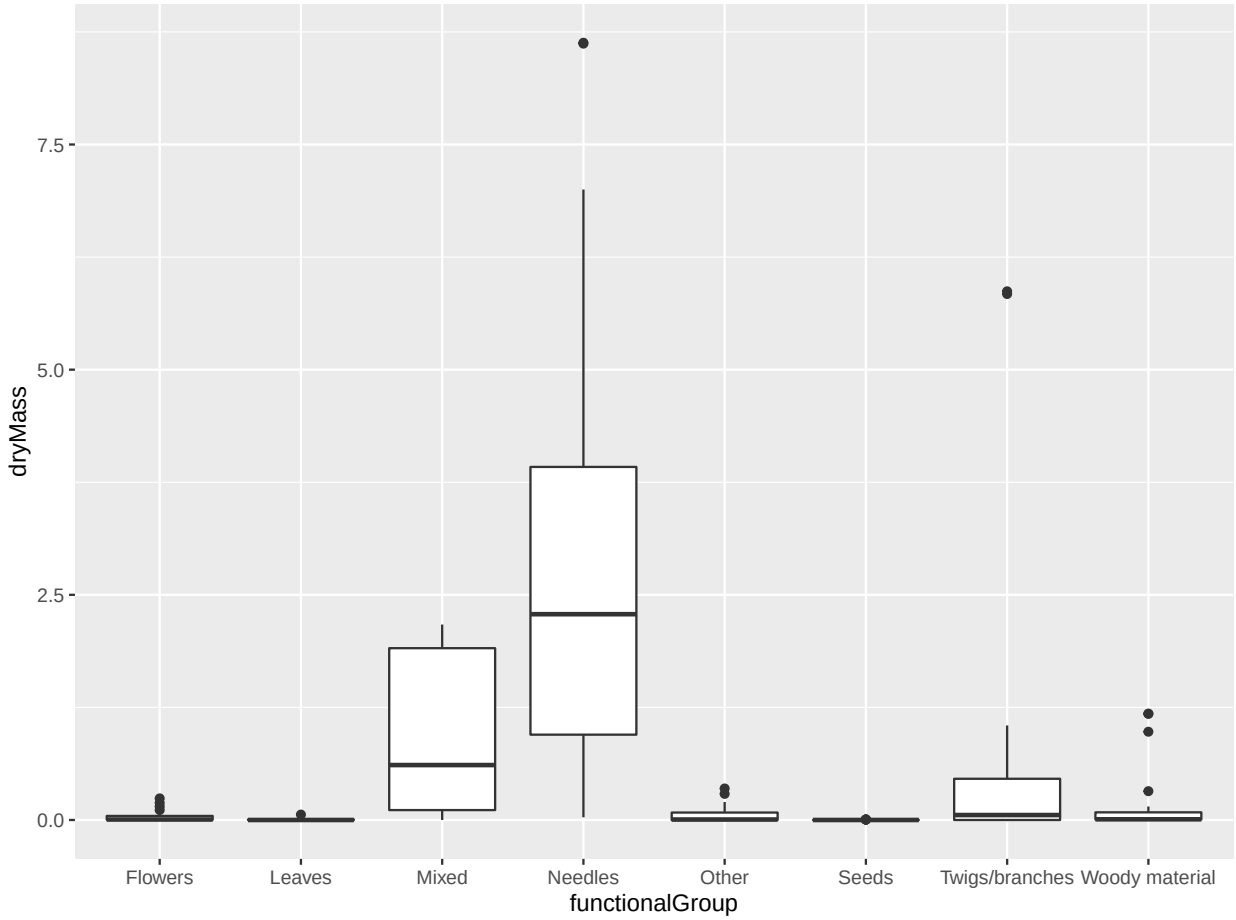
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

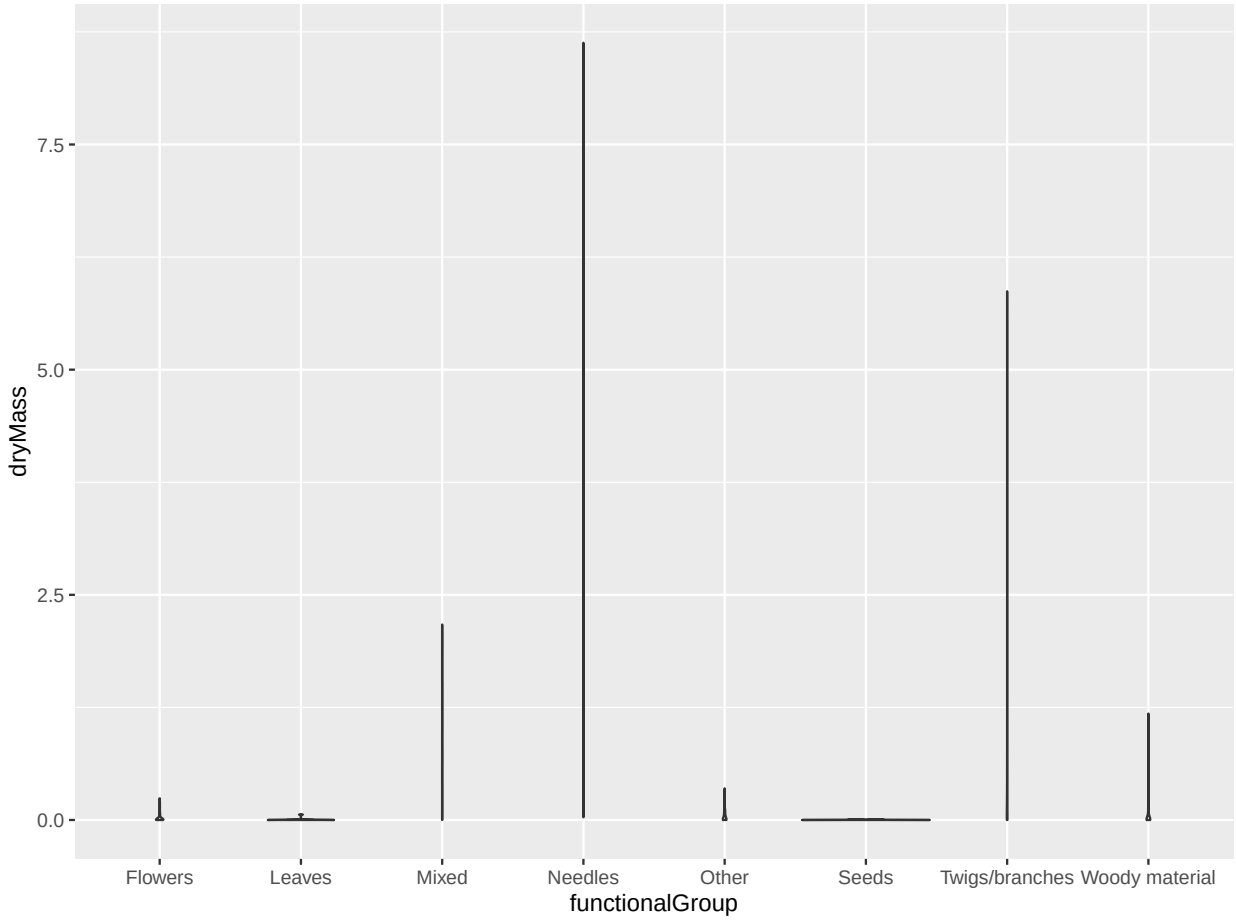


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




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



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

> Answer: Because the data like Needles is too dispersive, while the data like seeds is too concentrative. So using the violin plot, in this case, cannot show more characteristics of the distribution pattern than the boxplot.

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

> Answer: Needles.