Assignment 3: Data Exploration

Andrew Barfield

Fall 2024

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 Canvas.

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. Load necessary packages (tidyverse, lubridate, here), check your current working directory 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.

```r
# Loading packages
library(tidyverse)
library(here)
library(lubridate)

# Checking current working directory
getwd()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```r
here()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```r
# Uploading datasets
Neonics <- read.csv(file = here('./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv'),
    stringsAsFactors = TRUE)

Litter <- read.csv(file = here('./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: It is important to study ecotoxicology of neonicotinoids on insects because although the insecticides may be killing some pests which are harmful to agricultural crops, they might also have the unintended consquences of killing certain insects which are very beneficial to local ecosystems. Insecticides can harm endangered or threatened species, as well as species such as pollinators which ecosystems, and agrilcutural crops, heavily rely upon.

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: Litter and woody debris in forests provides many necessary ecosystem functions, including promoting biodiversity, playing a role in carbon and nutrient cycling, and providing habitat for aquatic and terrestrial ecosystems. In additon, litter and woody debris also provide structure for forest ecosystems that influence water and sediment flow.

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 woody debris are collected from elevated and ground traps, and are weighed and sorted into one of eight functional groups. 1. Sampling occurs in tower plots which are selected randomly within the 90% flux footprint of the primary and secondary airsheds. 2. In sites with forested tower airsheds, litter sampling takes place in twenty 40m x 40m plots. In sites with low-statured vegetation, litter sampling takes place in four 40m x 40m tower plots, plus twenty-six 20m x 20m plots. 3. One litter trap pair (one elevated and one ground) is deployed for every 400 m^2 plot area, resulting in 1-4 trap pairs per plot. These trap placements may either be targeted or randomized, depending on the vegetation.

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 on the "Effect" column, determine the most common effects that are studied. Why might these effects specifically be of interest? [Tip: The `sort()` command is useful for listing the values in order of magnitude. . .]

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

```
##       Hormone(s)        Histology       Physiology          Cell(s)
##                1                5                7                9
##     Biochemistry     Accumulation     Intoxication    Immunological
##               11               12               12               16
##       Morphology           Growth        Enzyme(s)         Genetics
##               22               38               62               82
##        Avoidance      Development     Reproduction Feeding behavior
##              102              136              197              255
##         Behavior        Mortality       Population
##              360             1493             1803
```

   Answer: The most common effects studied are population and mortality. These effects may specifically be of interest because scientists want to know if insecticides (specifically neonicotinoids in this case) are killing certain species, and what populations may potentially be most effected.

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: Explore the help on the `summary()` function, in particular the `maxsum` argument. . .]

```
summary(Neonics$Species.Common.Name, maxsum = 7)
```

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

   Answer: All of these species are flying insects, with 5/6 of the species being types of bees. Bees may be of particular interest because they are pollinators that are essential to plant health and ecosystem function, including agricultural crops. Parastic wasps also play a vital role in agricultural production because they eat many harmful pests that are often a danger to crops.

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? [Tip: Viewing the dataframe may be helpful. . .]

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

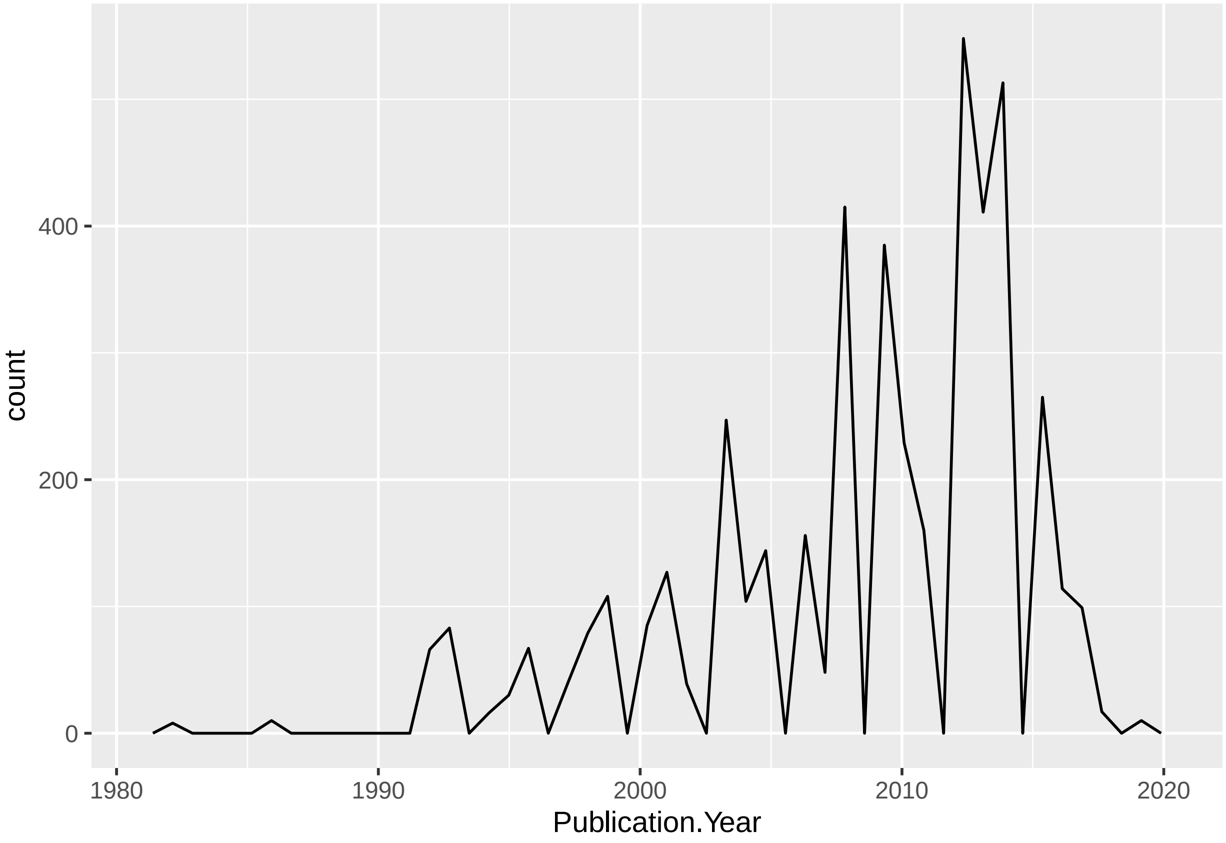
```
## [1] "factor"
```

> Answer: The ‘Conc.1.Author.’ column is classified as a factor. Even though concentrations are always a numeric value, these datapoints are classified as factors because they are assigned the labels of units of measurement. This means that different numeric values carry different meaning depending on what units are being used.

Explore your data graphically (Neonics)

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

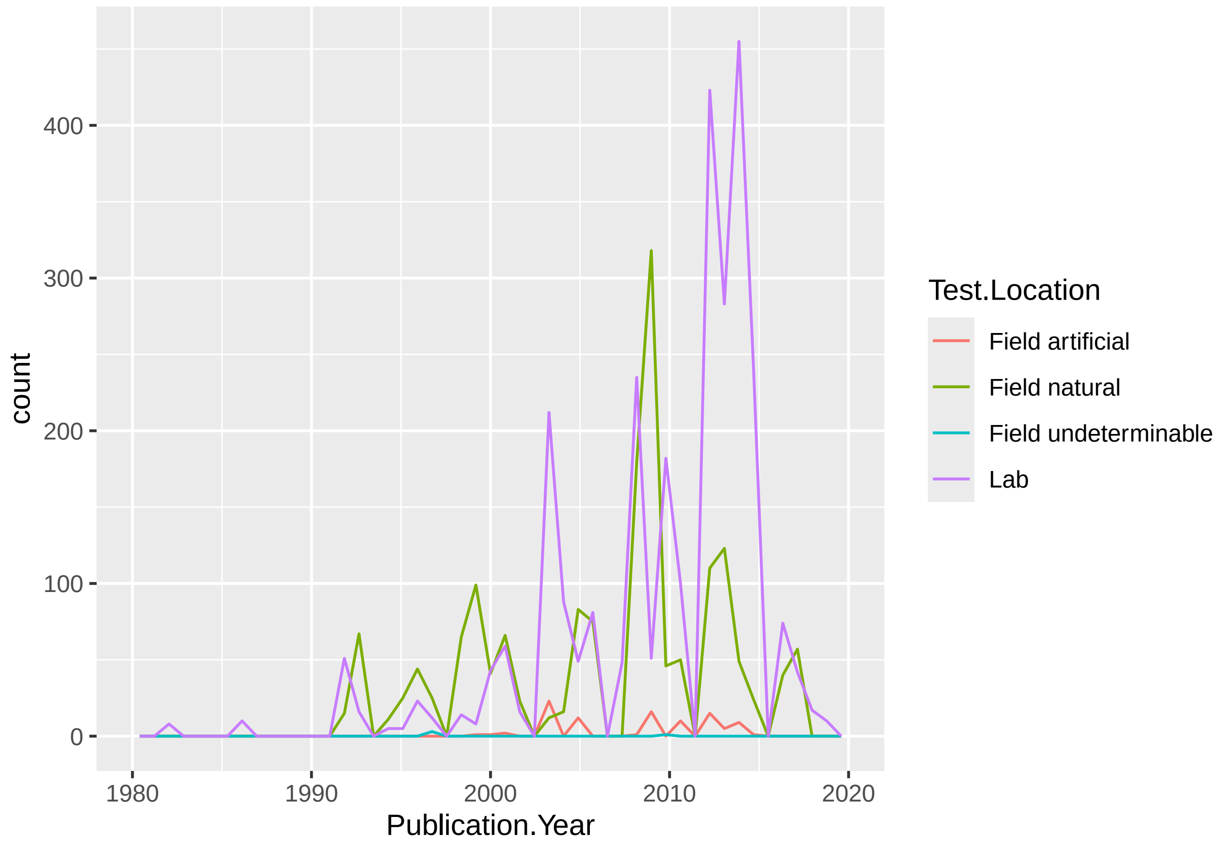
```
ggplot(Neonics) +
  geom_freqpoly(aes(x=Publication.Year), bins=50)
```



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 = 50) +
  scale_x_continuous(limits = c(1980,2020))
```

```
## Warning: Removed 8 rows containing missing values or values outside the scale range
## (‘geom_path()‘).
```



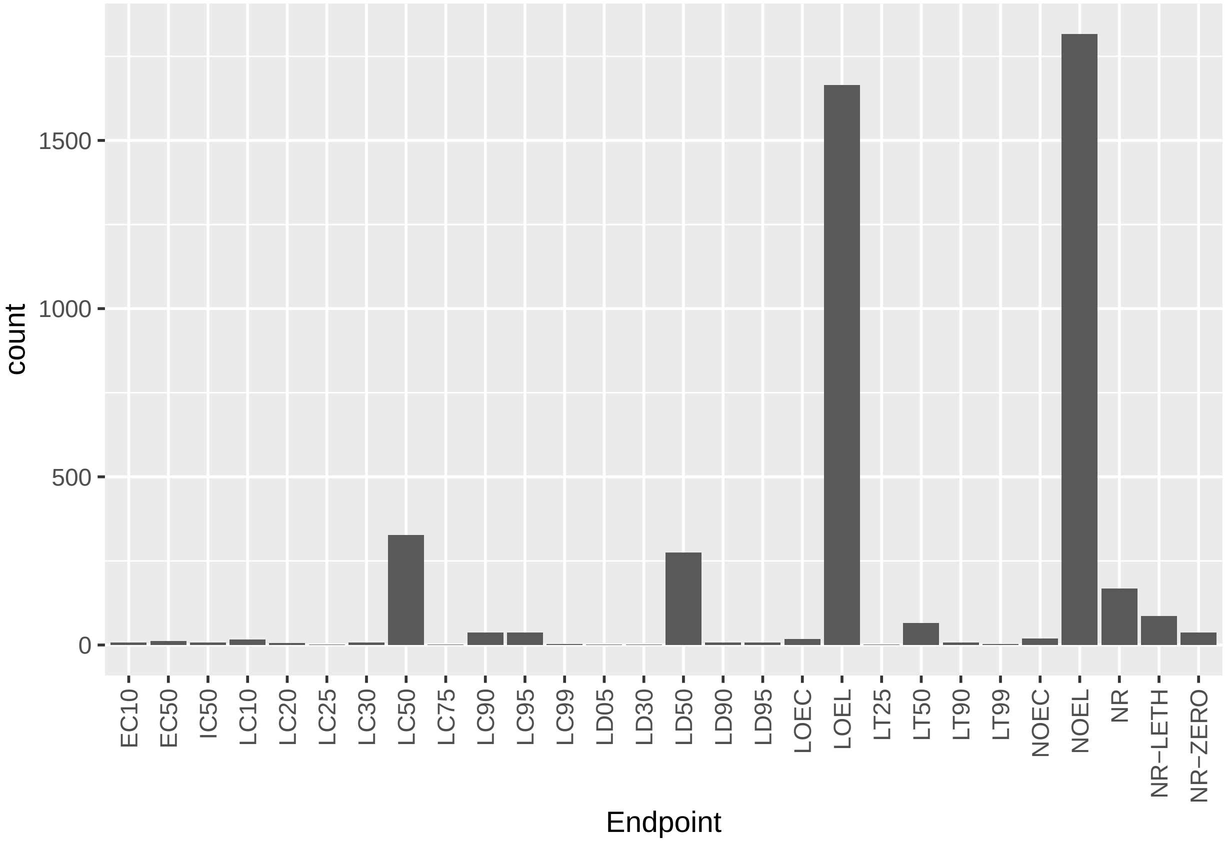
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are in the lab and in the natural field. It seems that over time, lab testing became more common and field testing decreased. In the earlier publication years, it seems that field testing was actually more common than lab testing (although not by much). Around the year 2000, lab testing seems to have taken off, although field testing has not completely gone away.

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



Answer: The most common end points are NOEL and LOEL. NOEL (No Observable Effect Level) is defined when the highest dose or concentration produces effects not significantly different from responses of controls. LOEL (Lowest Observable Effect Level) is defined when the lowest dose or concentration produces effects that are significantly different 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$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 different 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
```

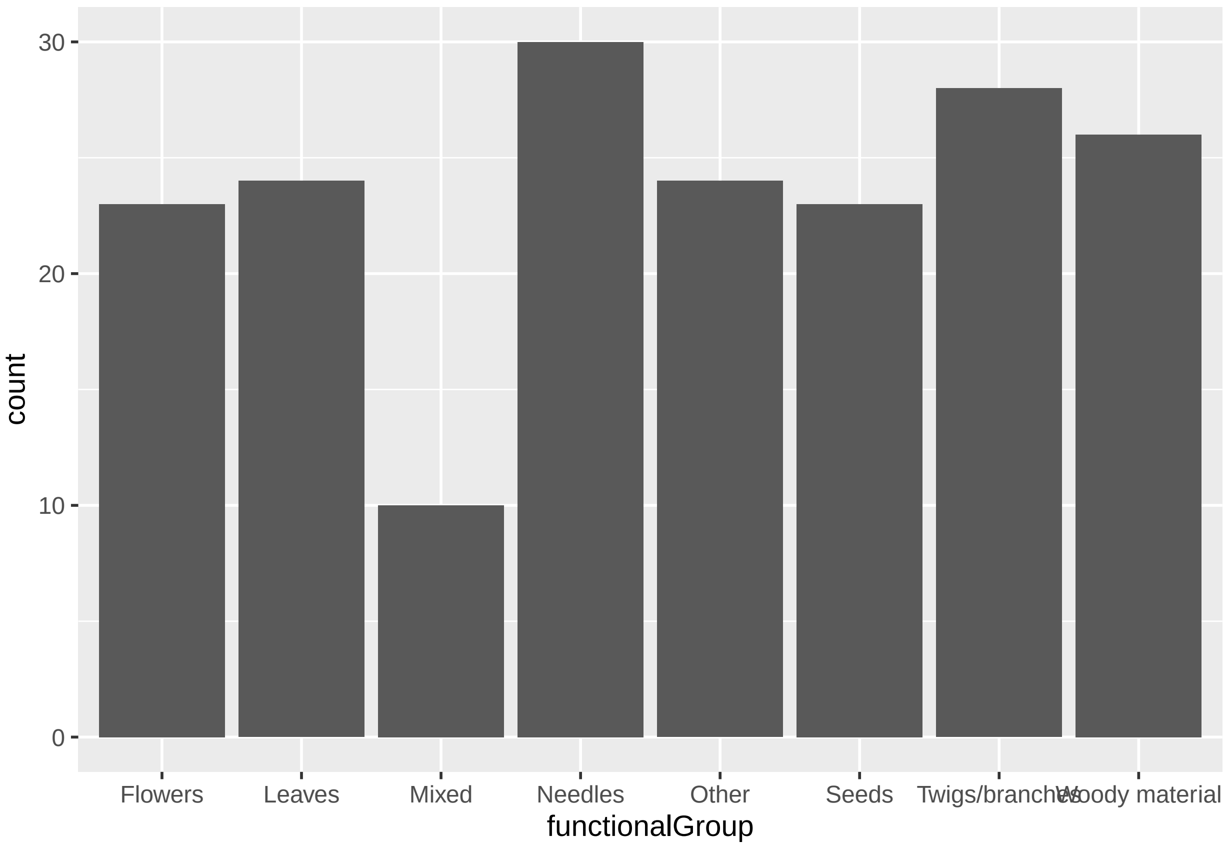
```
summary(Litter$plotID)
```

```
## 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 only tells you how many different plots were sampled through their individual ID numbers. The summary function, however, tells you both the different plots sampled (still by ID number) and the totals of each plot contained in your dataset.

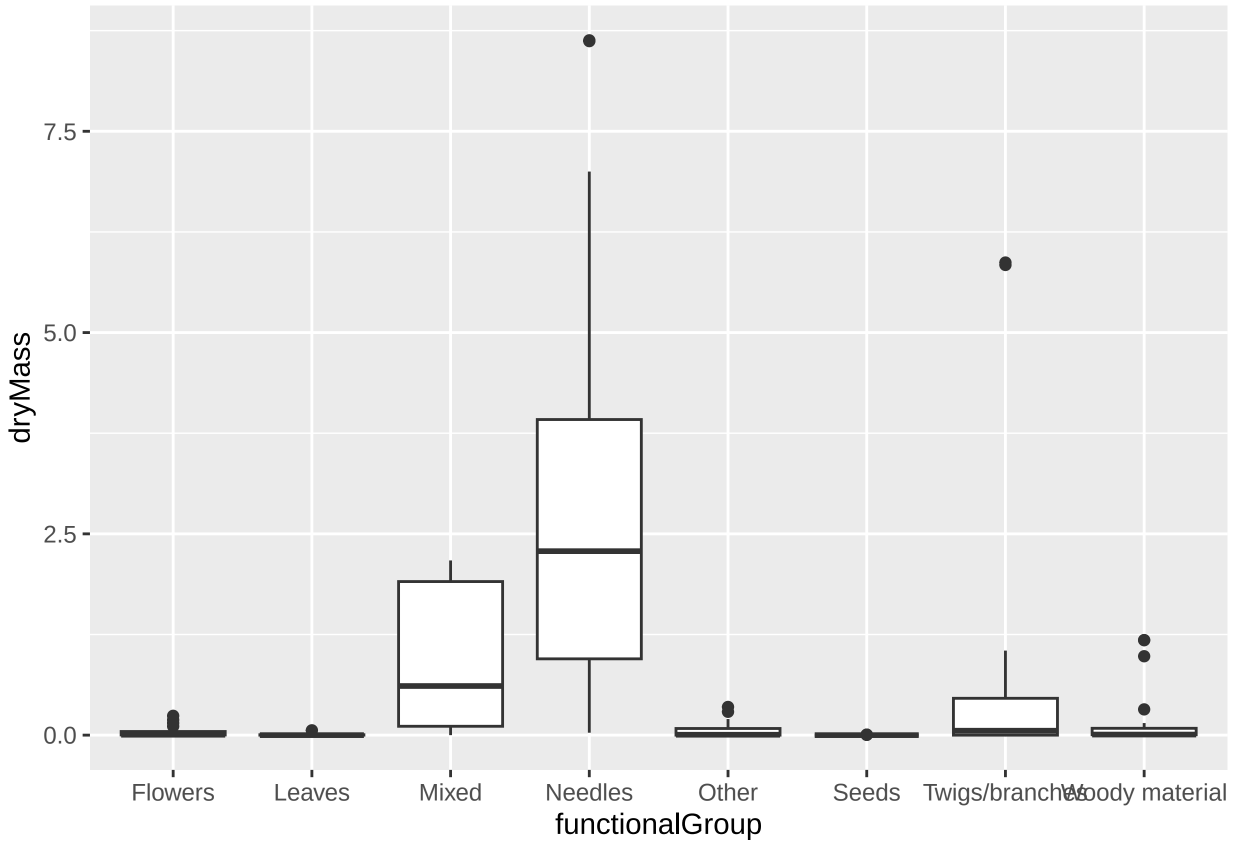
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

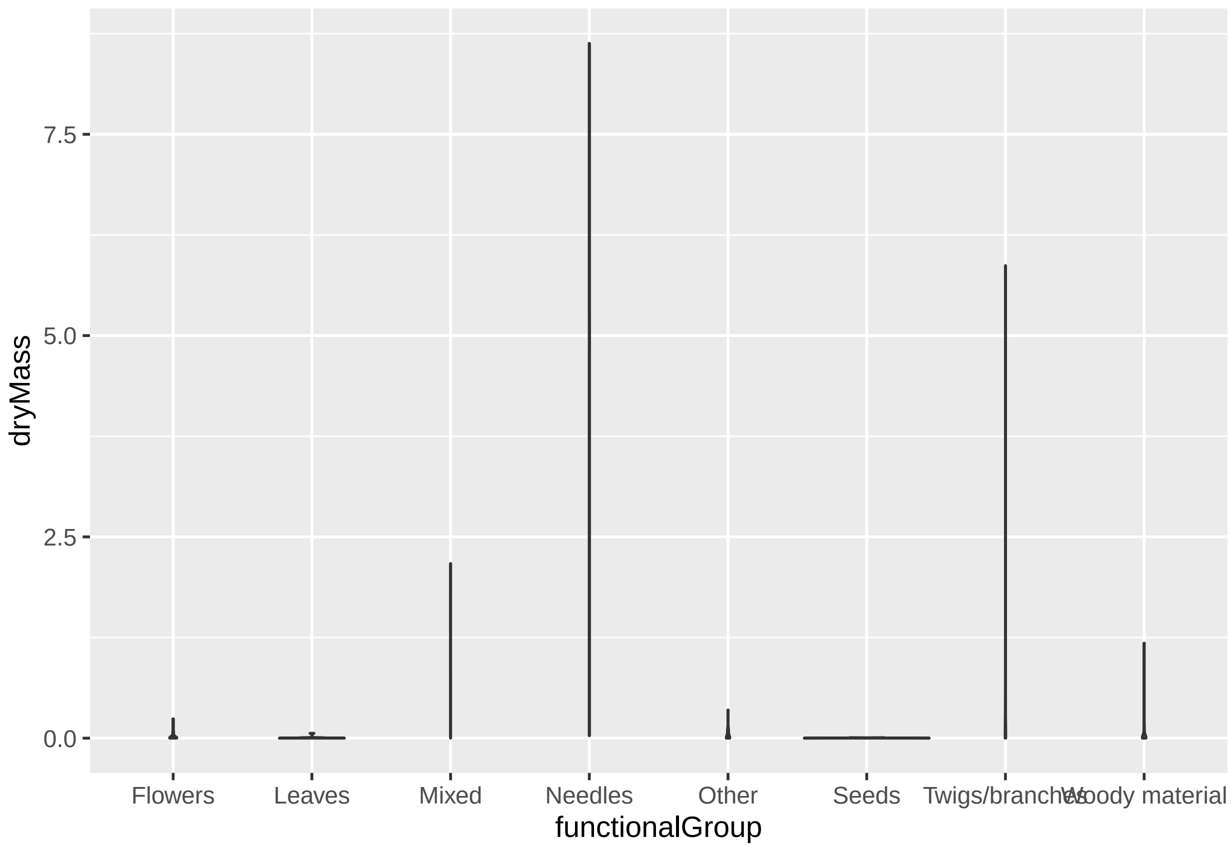


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
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: A boxplot is a more effective visualization option in this case because our data does not have very good distributions. Most of our data contains dry mass values that are very tiny (close to 0) and have little differentiation, making the violin plot not as useful as standard box plot.

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

> Answer: Needles and mixed, unsorted materials tend to have the highest biomass at these sites.