# Assignment 3: Data Exploration

Addie Navarro, Tuesday 8:30am Section #02

## OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Exploration.

### Directions

1. Change "Student Name, Section #" on line 3 (above) with your name and section number.
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., "FirstLast_A03_DataExploration.Rmd") prior to submission.

The completed exercise is due on January 25, 2022.

### 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 add the `stringsAsFactors = TRUE` parameter to the function when reading in the CSV files.**

```
getwd()
```

```
## [1] "Z:/EDA/Environmental_Data_Analytics_2022/Assignments"
```

```
#install.packages("tidyverse")
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 ecotoxicologoy of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: Neonicotinoids are a class of water soluble insectides that are intended to be applied to the soil and taken up by the plant to target invertebrate insects such as aphids and leave bees and other beneficial insects unharmed. However, research is showing that the neonicotinoids may be toxic to bees, not killing them directly, but contaminating flowers and nectar in low doses that cause the bees become disoriented and may be causing bee colony collapse disorder.

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 are important factors in a healthy forest ecosystem. Litter and woody debris are crucial sources of organic matter and influence healthy soil microorganisms, soil moisture, and soil temperature. This is important as climate change alters the micro-habitats of certain trees that rely on deep litter and woody debris to germinate.

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: * There are 1-4 litter trap pairs (one elevated and one ground trap) per plot * Sampling is conducted at NEON sites where woody vegetation is greater than 2 meters tall * Location of tower plots is random and traps are placed within the plots either randomly or targeted based 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?

```
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: These effects may be of interest to see how the chemical insecticide affects different species of insects. They're able to see how the insecticide affects the abundance or the mortality of the population, or if it affects the behavior or reproduction of the species, among other effects. Mostly it looks like the study affected mortality and population.

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, 7)
```

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

```
##                (Other)
##                   3083
```

Answer: These are all beneficial garden insects and are of more interest than other insects because of their role in pollination and keeping other populations of harmful insects at bay. The study is seeking to find out of the insecticide is harming beneficial insects.

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.Units..Author.)
```

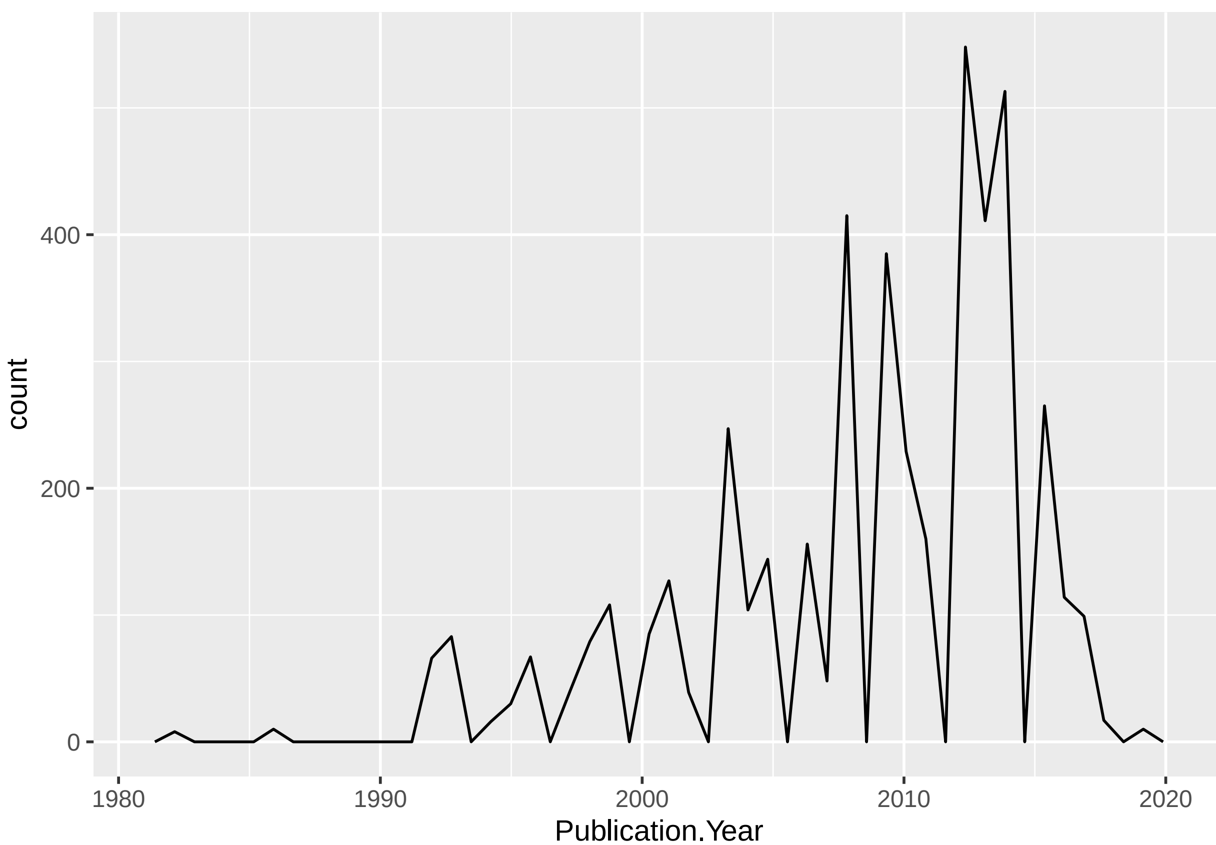
```
## [1] "factor"
```

Answer: The class of Conc.1..Author is a factor in the dataset and it's not numeric because it's categorical and therefore there are discrete values versus numeric values can be infinite.

## Explore your data graphically (Neonics)

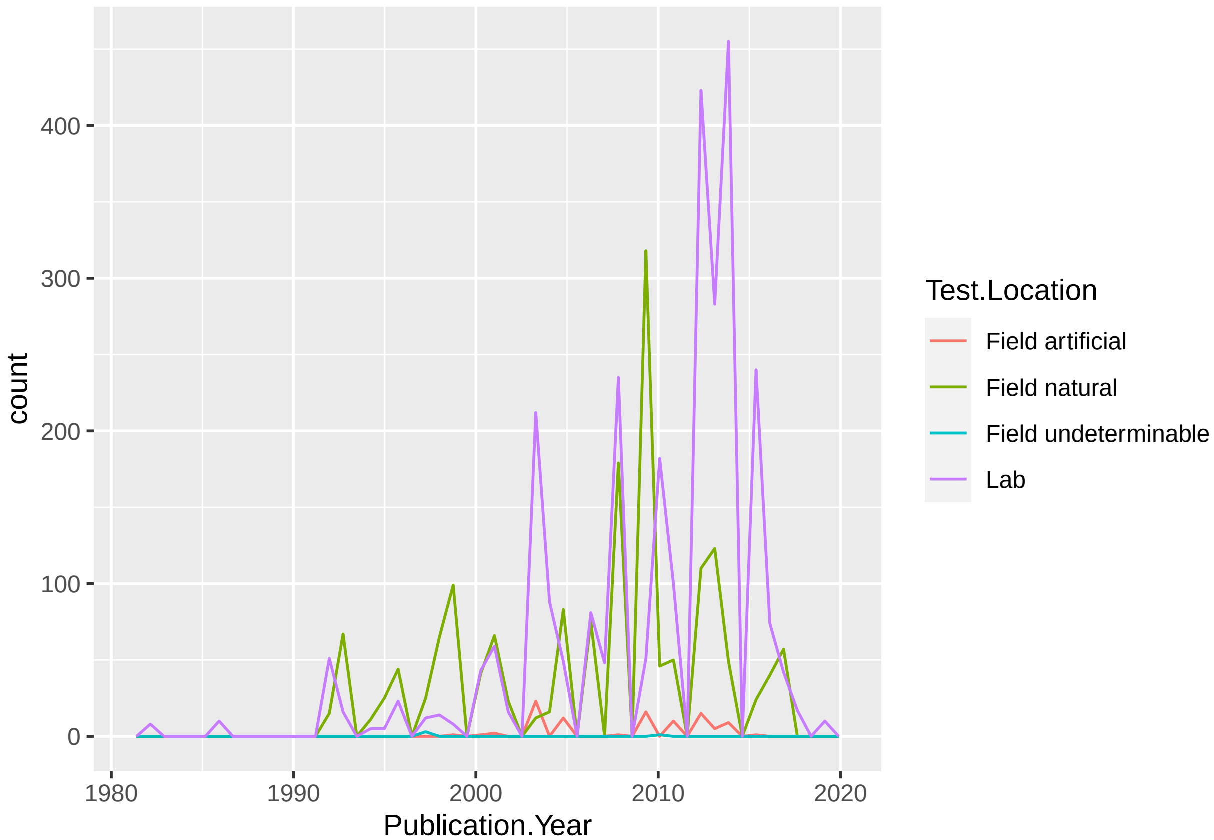
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
library(ggplot2)
library(dplyr)
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)
```




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

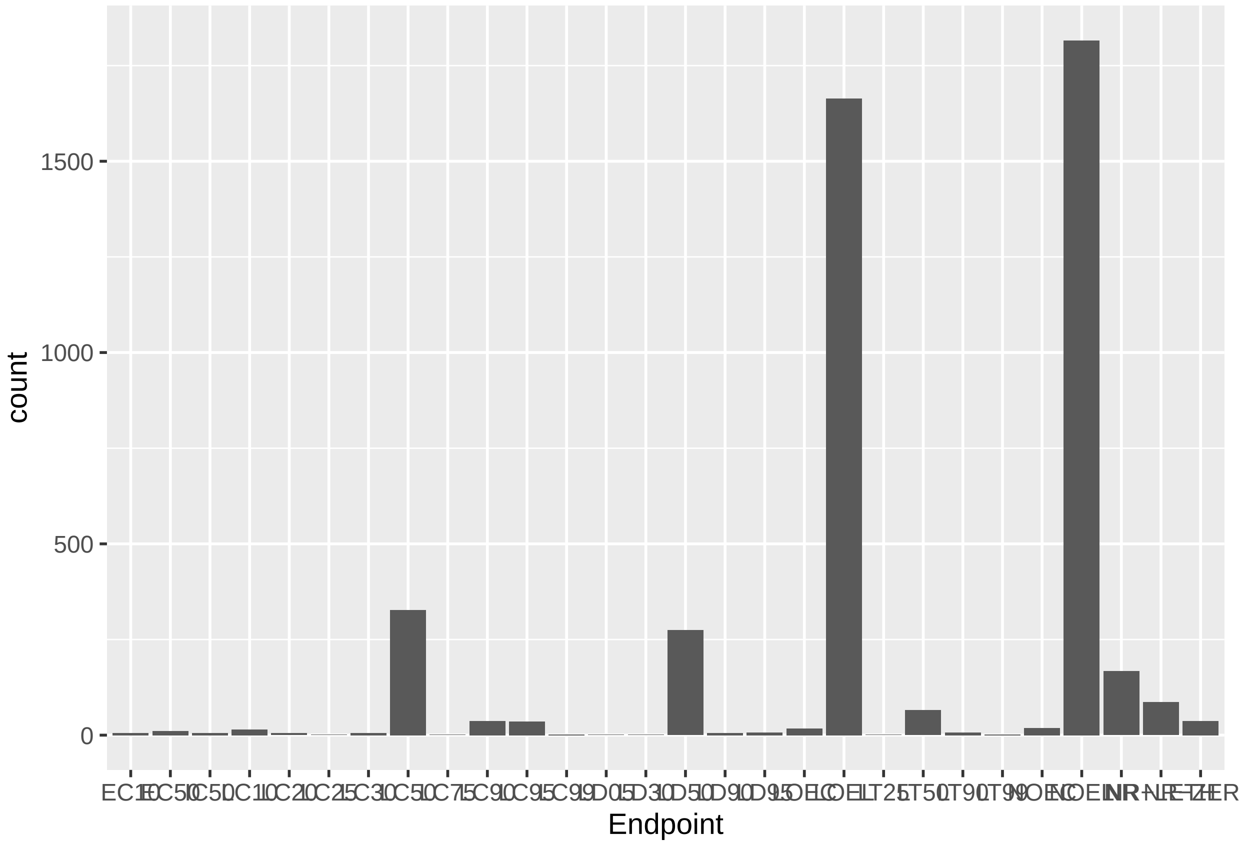
> Answer: The lab looks like the most common test location since the early 2000s, with natural field locations as a close second. Natural field testing locations looked more prominent in the 90s.

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.

```
summary(Neonics$Endpoint)
```

```
##     EC10     EC50     IC50     LC10     LC20     LC25     LC30     LC50     LC75     LC90
##        6       11        6       15        5        1        6      327        1       37
##     LC95     LC99     LD05     LD30     LD50     LD90     LD95     LOEC     LOEL     LT25
##       36        2        1        1      274        6        7       17     1664        1
##     LT50     LT90     LT99     NOEC     NOEL       NR  NR-LETH  NR-ZERO
##       65        7        2       19     1816      167       86       37
```

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




Answer: The two most common Endpoints are NOEL and LOEL. NOEL is defined as No Observable Effect Level and LOEL is defined as Lowest Observable Effect Level.

## 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.

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
Litter<-read.csv("../Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv", stringsAsFactors=TRUE)
class(Litter$collectDate) #it is not a date, it is a factor
```

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

```
Litter$collectDate<-as.Date(Litter$collectDate, format = "%Y-%m-%d")
class(Litter$collectDate)#now it's a date!
```

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

```
unique(Litter$collectDate, incomparables = FALSE)#collected on August 2, 2018 or August 30, 2018
```

```
## [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`?

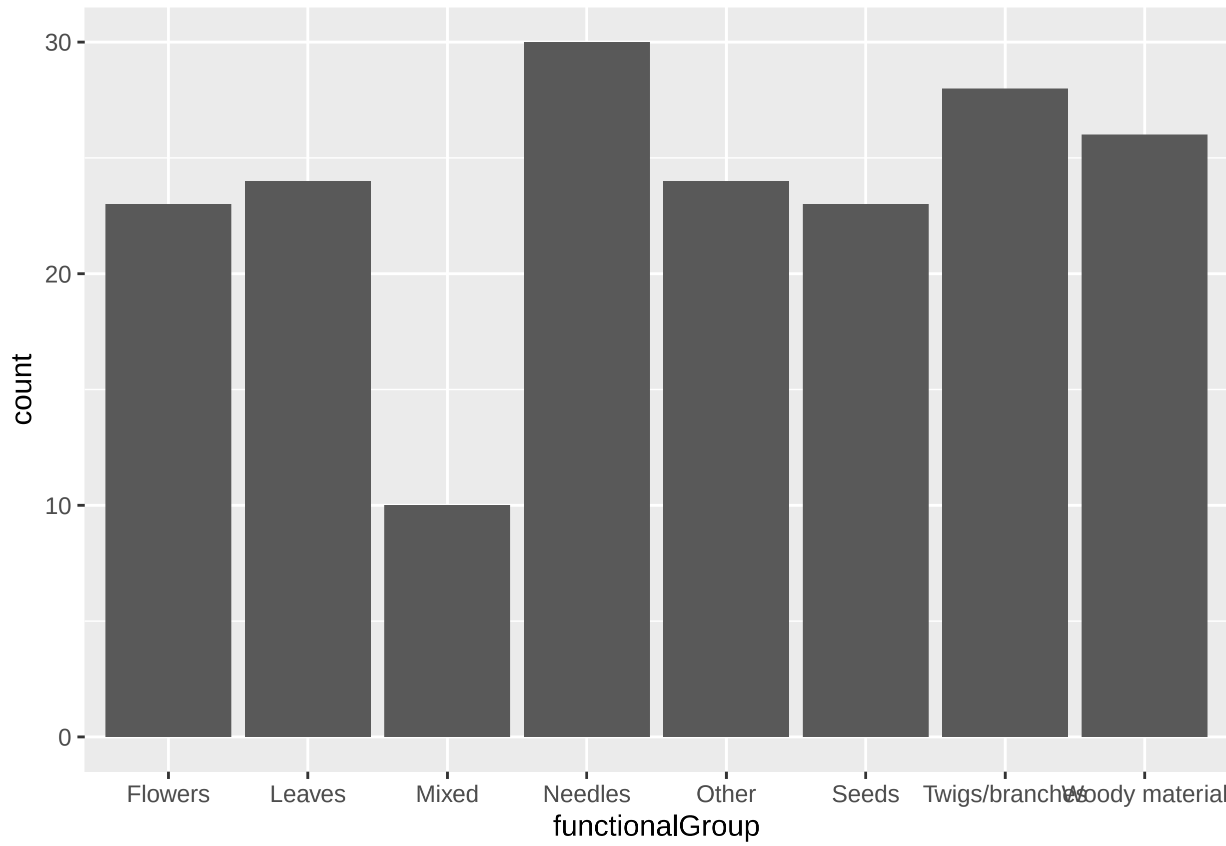
```
unique(Litter$plotID, incomparables = FALSE)
```

```
##  [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
```

Answer: There are 12 plots sampled at Niwot Ridge. This is different from summary because in summary, each of the plots are listed with the frequency number of samples at each plot versus in unique, it's just showing the number of unique plots that were sampled at Niwot Ridge.

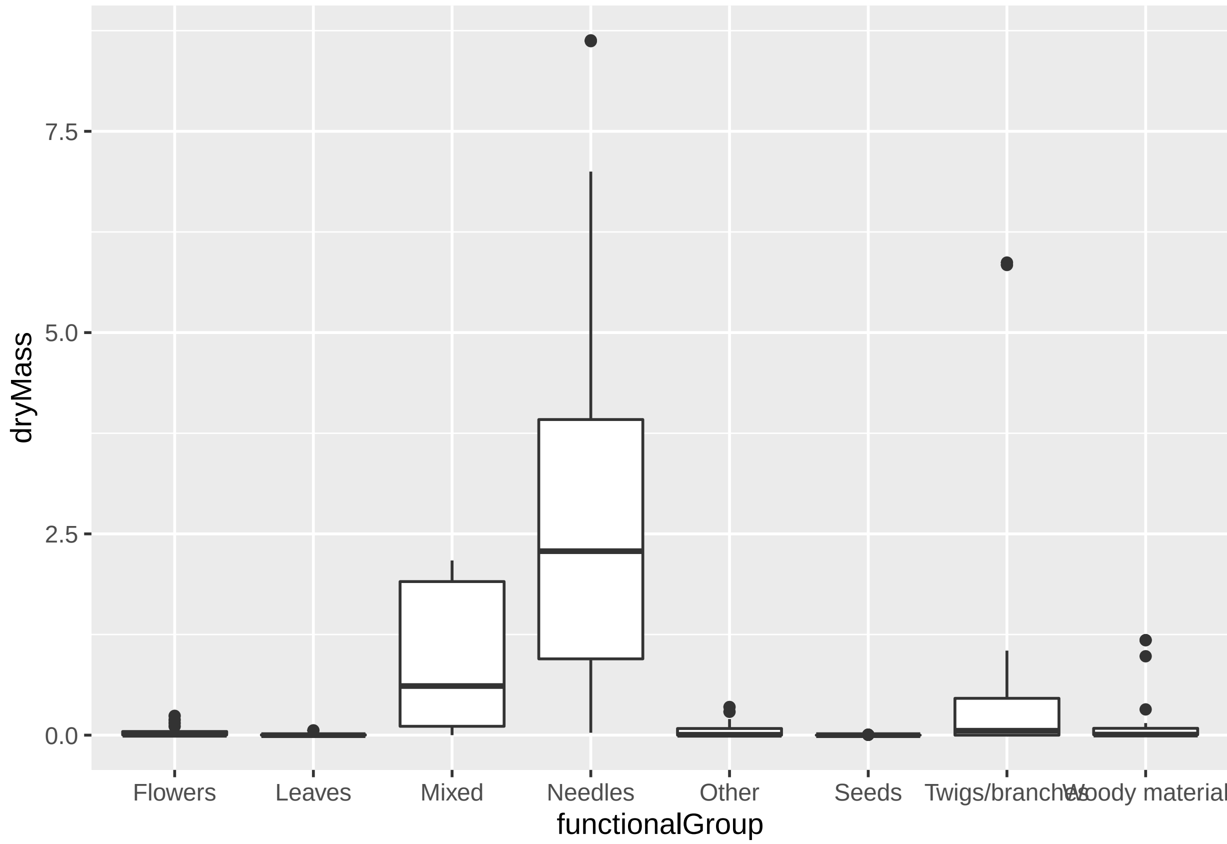
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

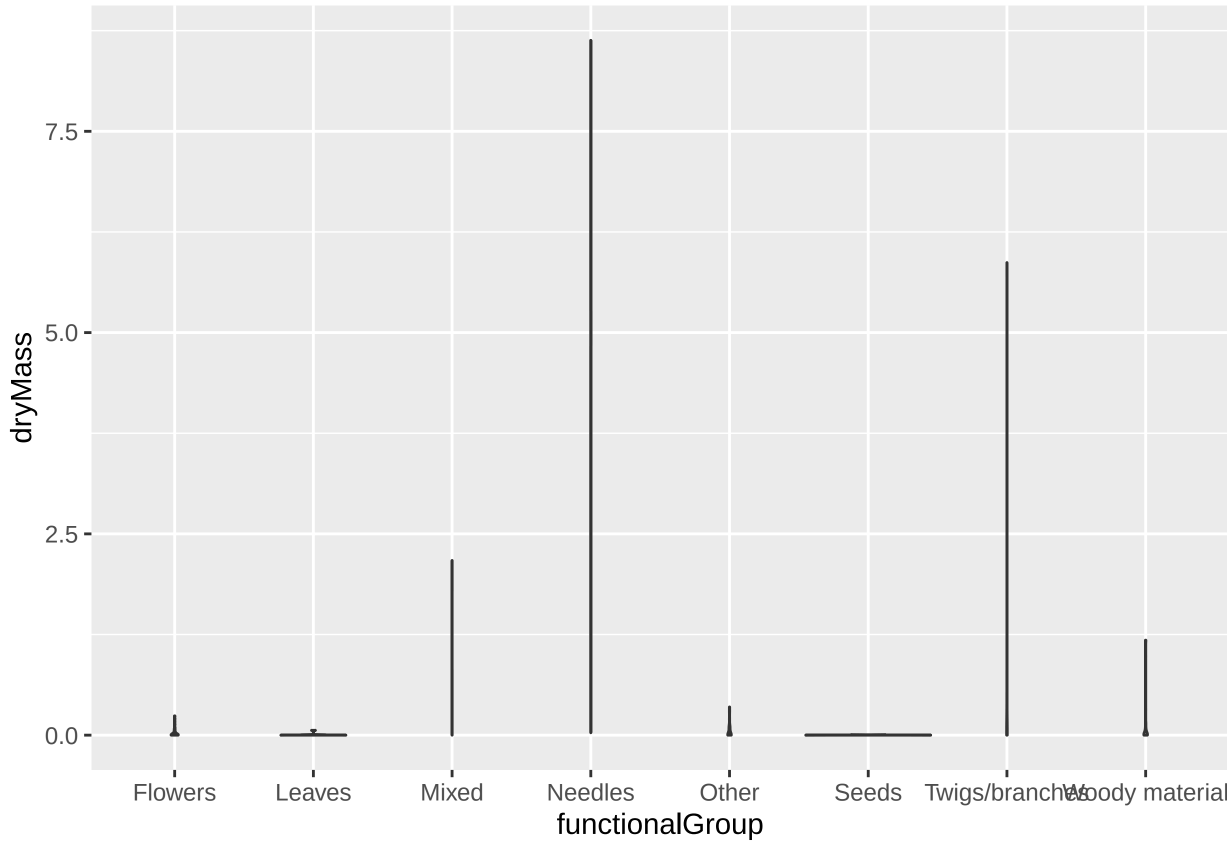



15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
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: The violin plot doesn't have enough data points of dry mass in each of the categories to show the peaks and valleys of the data. The boxplot shows more dimensions of the data in this case as we're interested in dry mass and can easily see the IQR of the different categories.

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

> Answer: Looking at the boxplot, it appears that needles have the highest biomass at these sites.