# Assignment 3: Data Exploration

Elsie Liu, Section #3

## 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 <>.

### Set up your R session

1. load necessary packages (tidyverse), upload: Neonics: ECOTOX neonicotinoid dataset (ECOTOX_Neonicotinoids_Insects_raw.csv) Litter: Niwot Ridge NEON dataset for litter and woody debris: (NEON_NIWO_Litter_massdata_2018-08_raw.csv)

```
library(tidyverse)
library(ggplot2)
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**: Neonicotinoid is a kind of active ingredient for insecticides and have been applied to control pests in a variety of crops; but at the same time, they may also have effect on non-target organisms such as pollinators (e.g. bees).

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**: The woody debris and litter could benefit forests' neutrient cycling, form habitat for some reptile species, or even become shelters for fishand amphibians while at the same time they could be an important feul source for wildfires.

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. sampling in 20 40m x 40m plots (for sites with low-statured vegetation over the tower airsheds, liter sampled in 4 40m x 40m tower plots plus 26 20m x 20m plots;
2. plot edges are separated by a distance 150% of one edge of the plot;
3. temporal desgin: ground traps are sampled once per year.

## Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

```
dim(Litter)
```

```
## [1] 188  19
```

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**: "Population" is the most common effect in the "Effect" column. Researchers might be interested in how one (or several) effect could influence insects' behavior seperately (or together).

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.

```
Neonics.summ <- as.data.frame(summary(Neonics$Species.Common.Name))
colnames(Neonics.summ) <- "count"
Neonics.summ$species <- rownames(Neonics.summ)
head(Neonics.summ[order(Neonics.summ$count, decreasing = T),],n=7)
```

```
##                       count               species
## (Other)                 670               (Other)
## Honey Bee               667             Honey Bee
## Parasitic Wasp          285        Parasitic Wasp
## Buff Tailed Bumblebee   183 Buff Tailed Bumblebee
## Carniolan Honey Bee     152   Carniolan Honey Bee
## Bumble Bee              140            Bumble Bee
```

```
## Italian Honeybee         113       Italian Honeybee
```

**Answer**: Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, Italian Honeybee are the most coommly studied species. These insect species are polinators and play important roles in the ecosystem. Once these species have been influenced by the pasticide, some primary producer that rely on them would lost their way of reproduction.

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"
```

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

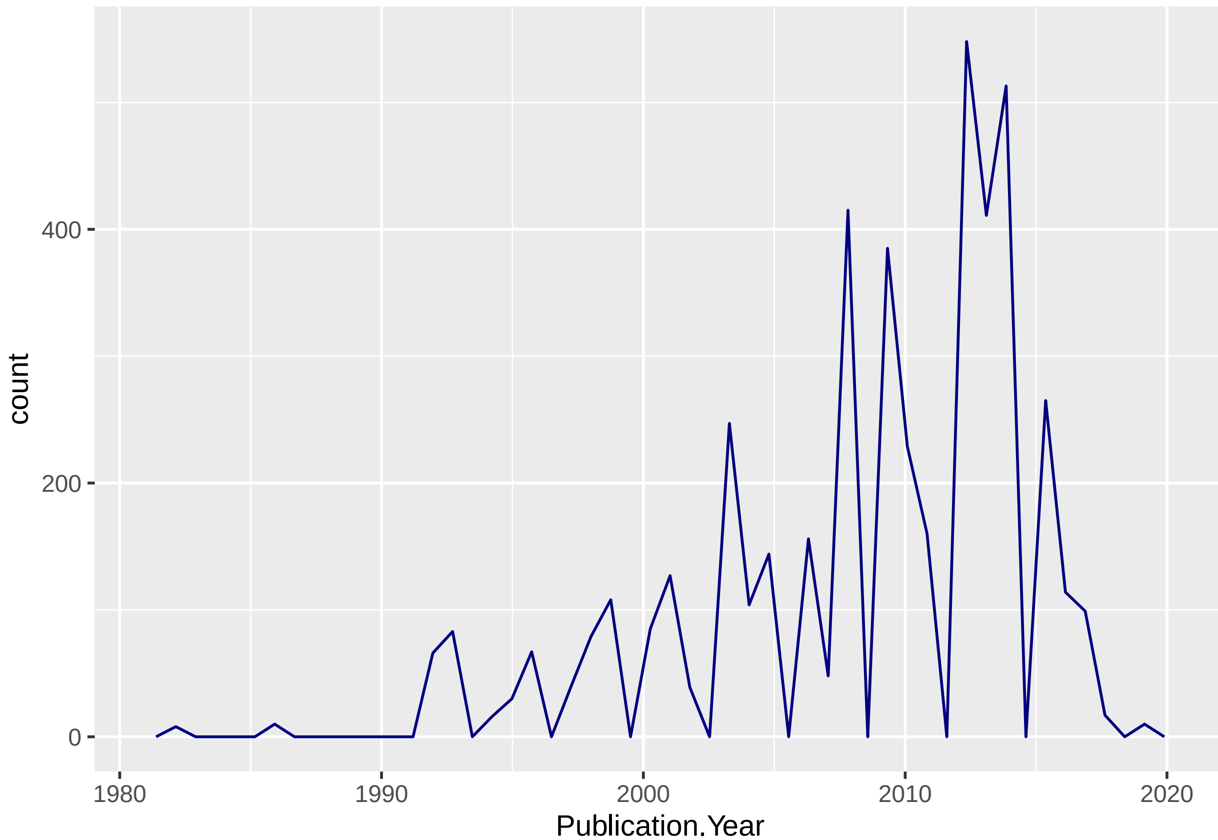
```
##    0.37/      10/      NR/       NR        1     1023    0.40/       2/
##      208      127      108       94       82       80       69       63
##       10   0.053/      100      50/     0.5/     0.03    0.05/     0.45
##       62       59       56       51       45       44       43       43
##     0.1/    0.45/     1.0/    2.27/       50    0.125     500/      0.5
##       42       40       40       40       36       33       33       32
##   0.048/    0.15/       1/       48    25.0/      12/    0.027      2.4
##       30       30       30       30       28       27       26       26
##     0.2/    0.56/     100/        3    0.01/    1000/       3/    0.336
##       25       24       23       23       22       22       22       21
##     1.5/     0.05      1.5    2.60/    20.0/        6    6.80/    62.5/
##       21       20       20       20       20       20       20       20
##    0.005     0.4/    0.18/     0.3/     1000       40 0.00355/      0.1
##       18       18       17       17       17       17       16       16
##      0.4     150/      300      80/    0.053     0.24     0.28     125/
##       16       16       16       16       15       15       15       15
##        9   0.0001  0.0004/   0.084/     0.15      0.6    12.5/   144.0/
##       15       14       14       14       14       14       14       14
##     350/    40.0/      48/       56      84/    0.17/      125       14
##       14       14       14       14       14       13       13       13
##       16       17   0.047/    0.25/    0.28/    1.28/    1.81/      112
##       13       13       12       12       12       12       12       12
##      150     2.5/       25      60/      75/    0.02/   0.025/     0.29
##       12       12       12       12       12       11       11       11
##    37.5/       4/        5  (Other)
##       11       11       11     1817
```

**Answer**: According to EPA's ECOTOX Web site, "Conc.1..Author." stands for "Author Reported and Standardized Concentration". The class of this field is factor but not numeric because there are some none numeric sting like "NR/" or "numbers+/".

## Explore your data graphically (Neonics)

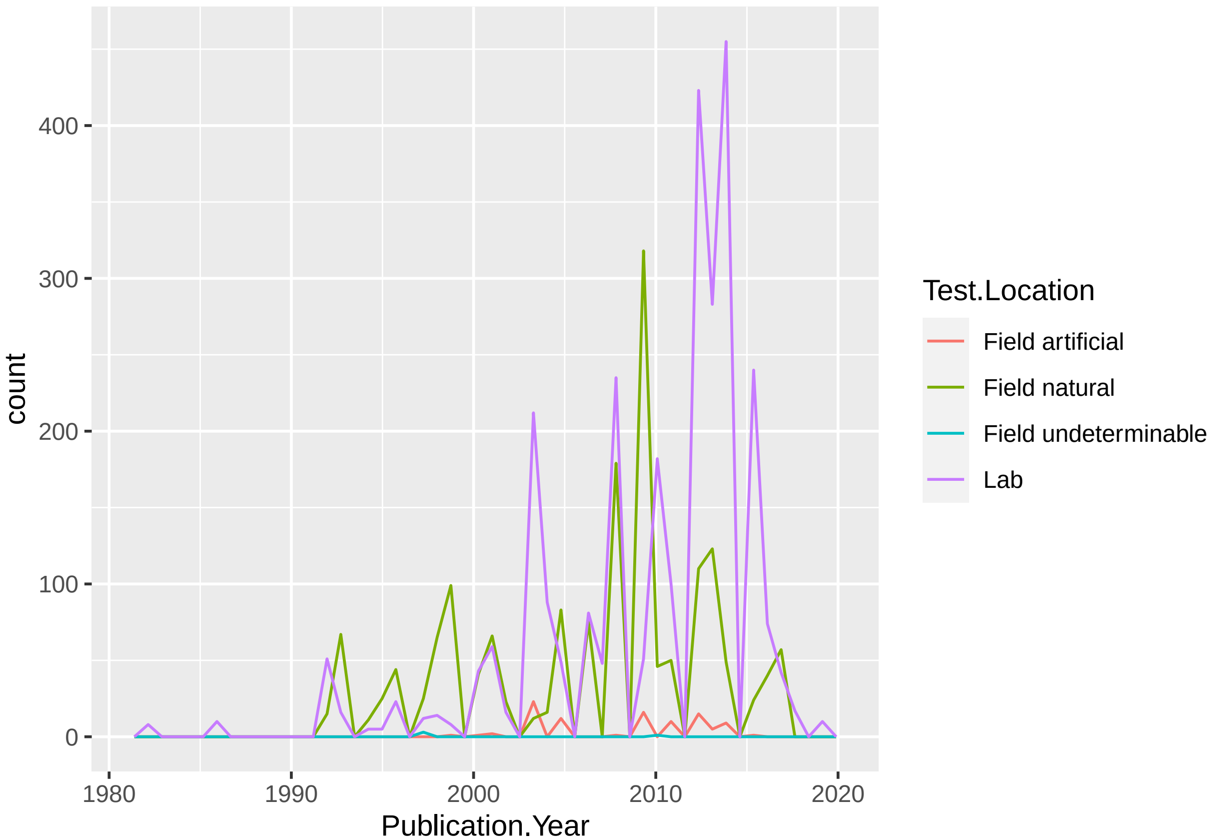
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

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




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)
```

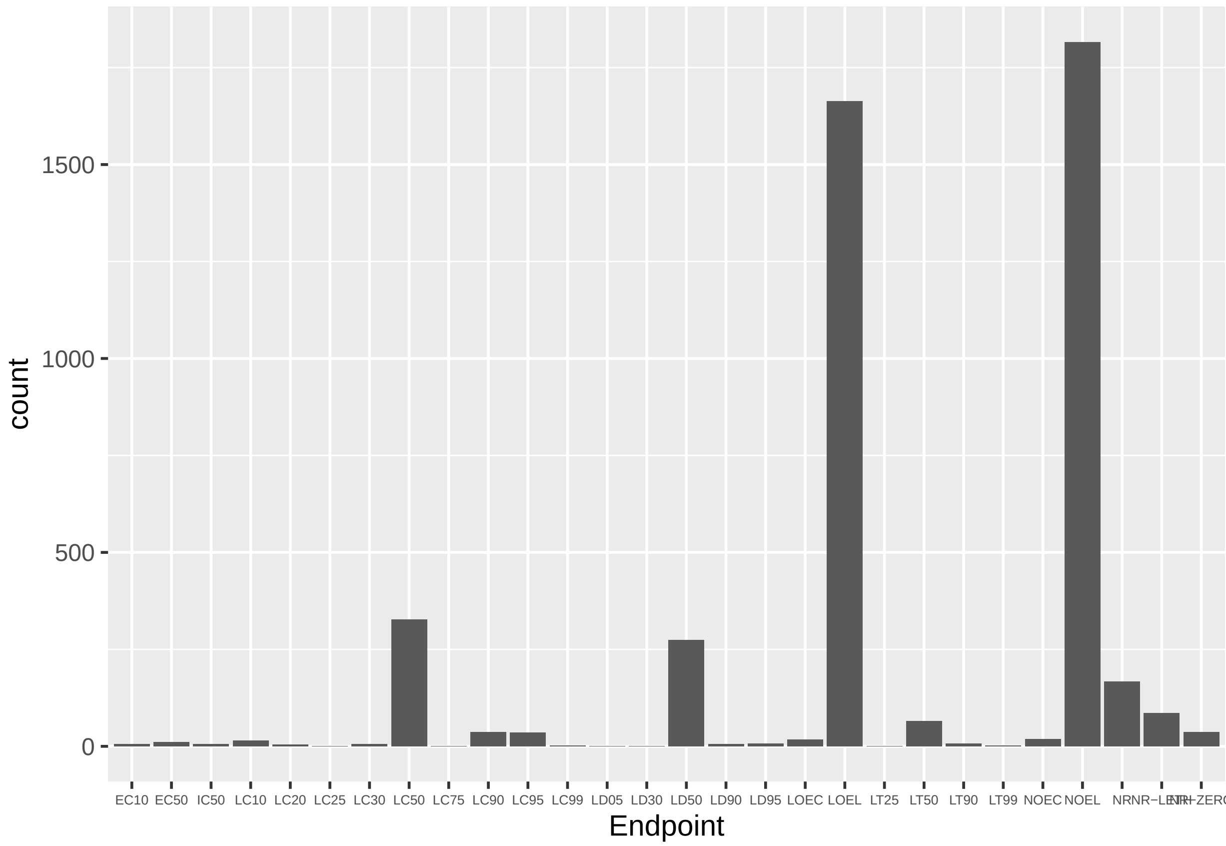



Interpret this graph. What are the most common test locations, and do they differ over time?

> **Answer**: According to the graph above, Lab/Field natural are most common test locations. "field natural" is more common in 1990s and 2009 and lab is more common after 2000 (several exceptions).

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, aes(x = Endpoint)) +
  geom_bar() +
  theme(axis.text.x = element_text(size = 5))
```




**Answer**: The two most common end points are NOEL (No-observable-effect-level) & LOEL (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
```

```
class(Litter$collectDate)
```

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

```
Litter$collectDate <- ymd(Litter$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`?

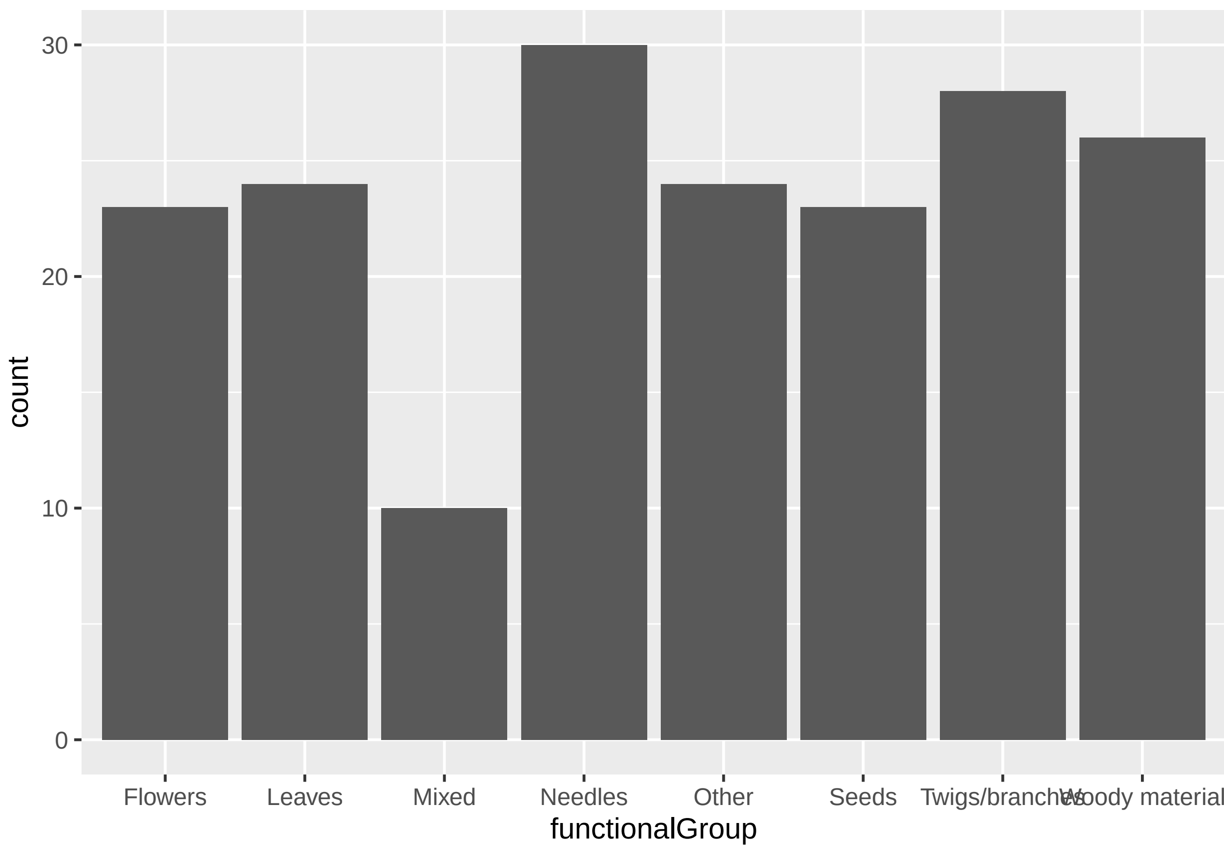
```
unique(Litter$plotID)
```

```
##  [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
```

> **Answer**: 12 plots were sampled at Niwot Ridge. "unique()" only counts levels while "summary()" also counts the frequency of these levels.

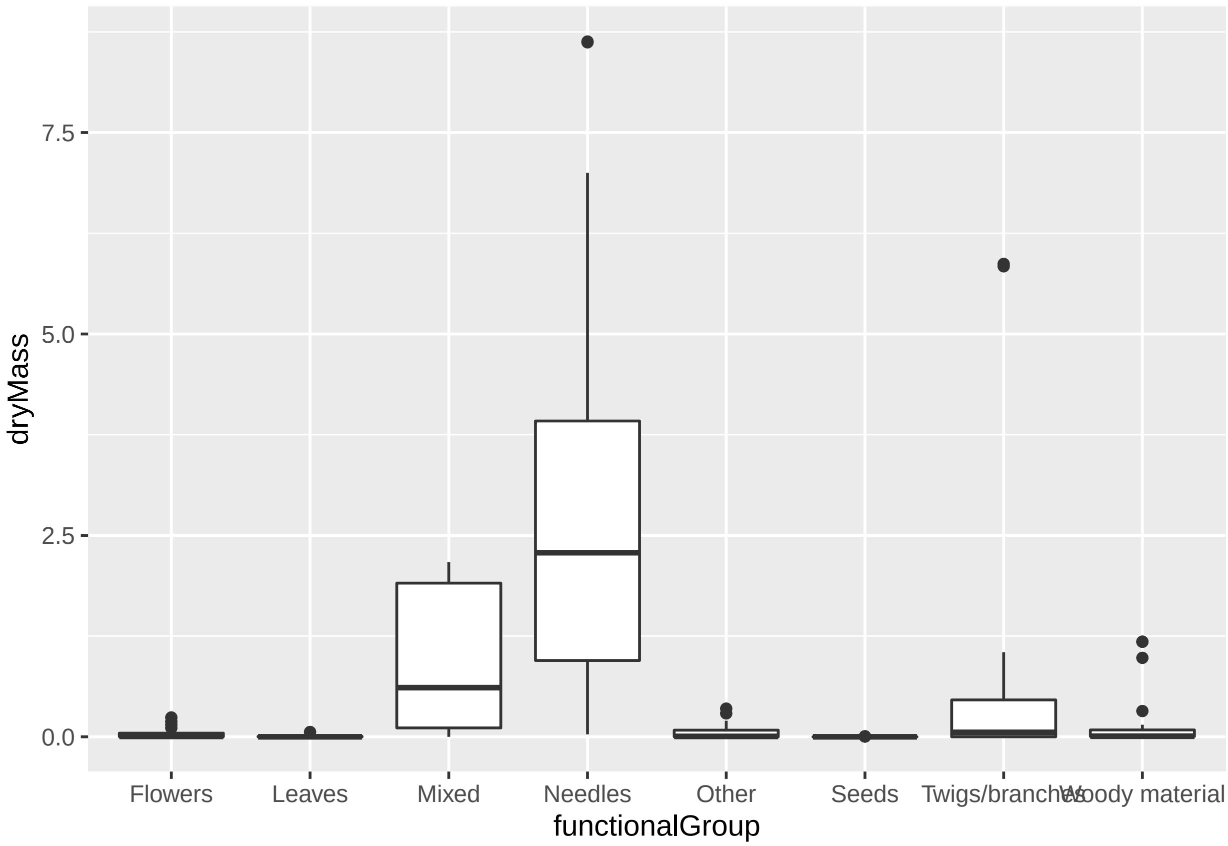
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

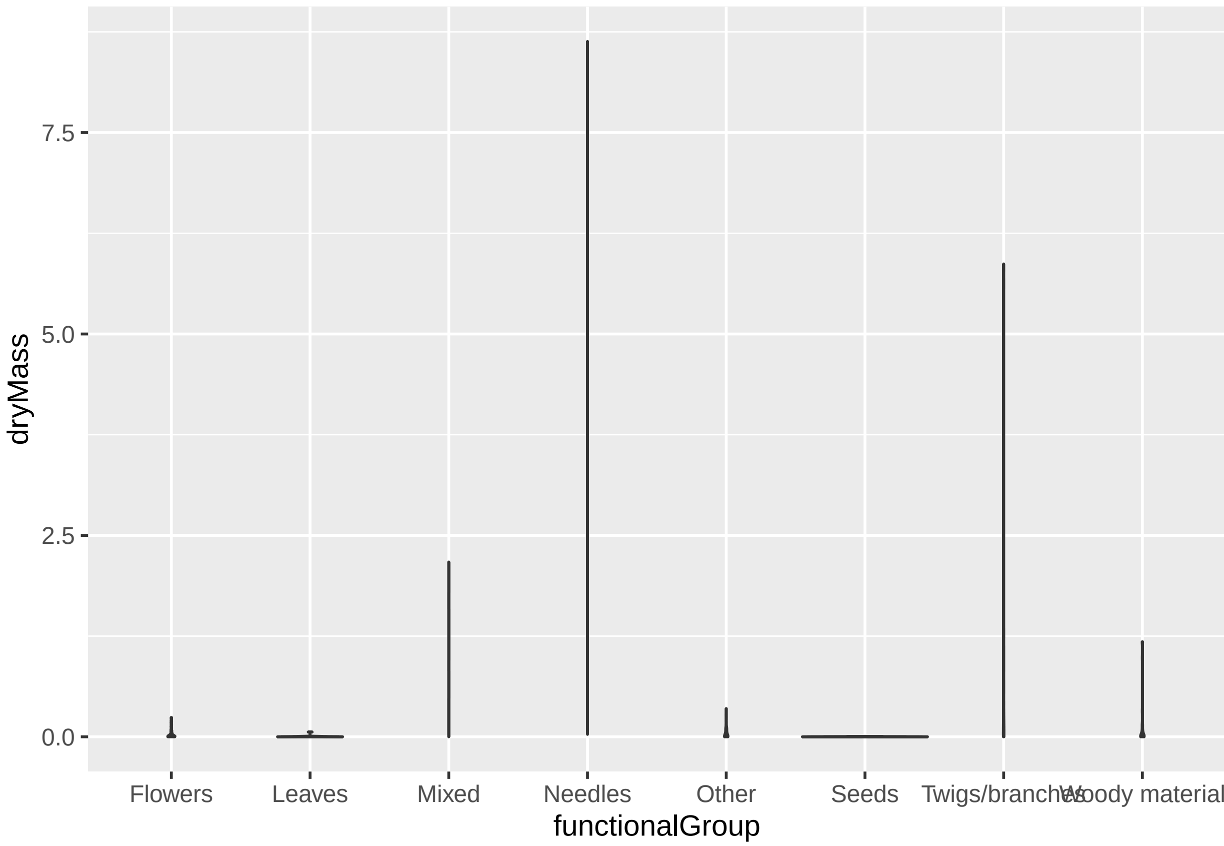



```
ggplot(data = Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75))
```

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

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

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




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

> **Answer**: The boxplot could clearly show the outliers and the area of box itself (representing the distribution of dry mass in each functional group). On the contrary, the violin plot failed to show the distribution of groups like Mixed, Needles and Twigs/branches as these groups have relatively large ranges and shrink to single lines.

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

> **Answer**: Needles are of the highest biomass.