Assignment 3: Data Exploration

Joshua Frear

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.

```
#check working directory
getwd()
```

```
## [1] "D:/Data_872/Environmental_Data_Analytics_2021/Assignments"
```

```
#load packages
#install.packages('tidyverse')
library(tidyverse)
#load ECOTOX dataset
Neonics <- read.csv("../Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv")
#load Niwot Ridge NEON dataset
Litter <- read.csv("../Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv")
```

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: Two concerns at the ecological level are that they will kill off beneficial insects like honeybees and/or lead to drops in bird populations (not enough insect prey remaining).

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: rates of decomp of litter and woody debris are key to rates of CO2 re-entering the atmosphere from sequestration in trees (Gora et al 2018). Litter and woody debris also provide habitat structure and resources for microorganisms and invertabrates (Zimmer, Encyclopedia of Ecology, 2019)

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: *Masses reported indicate one collection event from one trap.* Weights are accurate to 0.01g. Weights below that indicate presence. *the trapID is a unique code for each trap, with the last three digits indicating its clipCell placement within the plot, which can be deciphered using a NEON lookup table

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 "Effects" column, determine the most common effects that are studied. Why might these effects specifically be of interest?

```
#summary does not produce a useful result to determine how common effects are
#because it is a character column.
summary(Neonics$Effect)
```

```
##    Length     Class      Mode
##      4623 character character
```

```
#Using the count() function instead produces a quick summation
#of how frequent different effects are.
count(Neonics, Effect)
```

```
##              Effect    n
## 1      Accumulation   12
## 2         Avoidance  102
## 3          Behavior  360
## 4      Biochemistry   11
## 5           Cell(s)    9
## 6       Development  136
## 7         Enzyme(s)   62
## 8  Feeding behavior  255
## 9           Genetics   82
## 10            Growth   38
## 11        Histology    5
## 12       Hormone(s)    1
## 13    Immunological   16
## 14     Intoxication   12
## 15       Morphology   22
## 16        Mortality 1493
```

```
## 17       Physiology    7
## 18       Population 1803
## 19     Reproduction  197
```

Answer: Mortality and Population are the most common. Mortality makes sense - we are studying poisons and want to see if organism death results or not. Population is defined in the supplemental document on pg 94 as "measurements and endpoints relating to a group of organisms or plants of the same species occupying the same area at a given time." This description is somewhat vague, suggesting that a lot of different experiments could fall into this category.

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.

```
#Again, not useful.
summary(Neonics$Species.Common.Name)
```

```
##    Length     Class      Mode
##      4623 character character
```

```
#A simple count() function doesn't sort by most common.
#count(Neonics, Species.Common.Name)
#Set the count() to a variable, which can then be opened and sorted by n with a click
sp_freq <- count(Neonics, Species.Common.Name)
```

Answer: Honey Bee - 667; Parasitic Wasp - 285; Buff Tailed Bumblebee - 183; Carniolan Honey Bee - 152; Bumble Bee - 140; Italian Honeybee - 113.

These insects tend to be pollinators which are vital for ecosystems and agriculture. The exception, parasitic wasps, are often considered beneficial to humans because they parasitize pest 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..Author.)
```

```
## [1] "character"
```

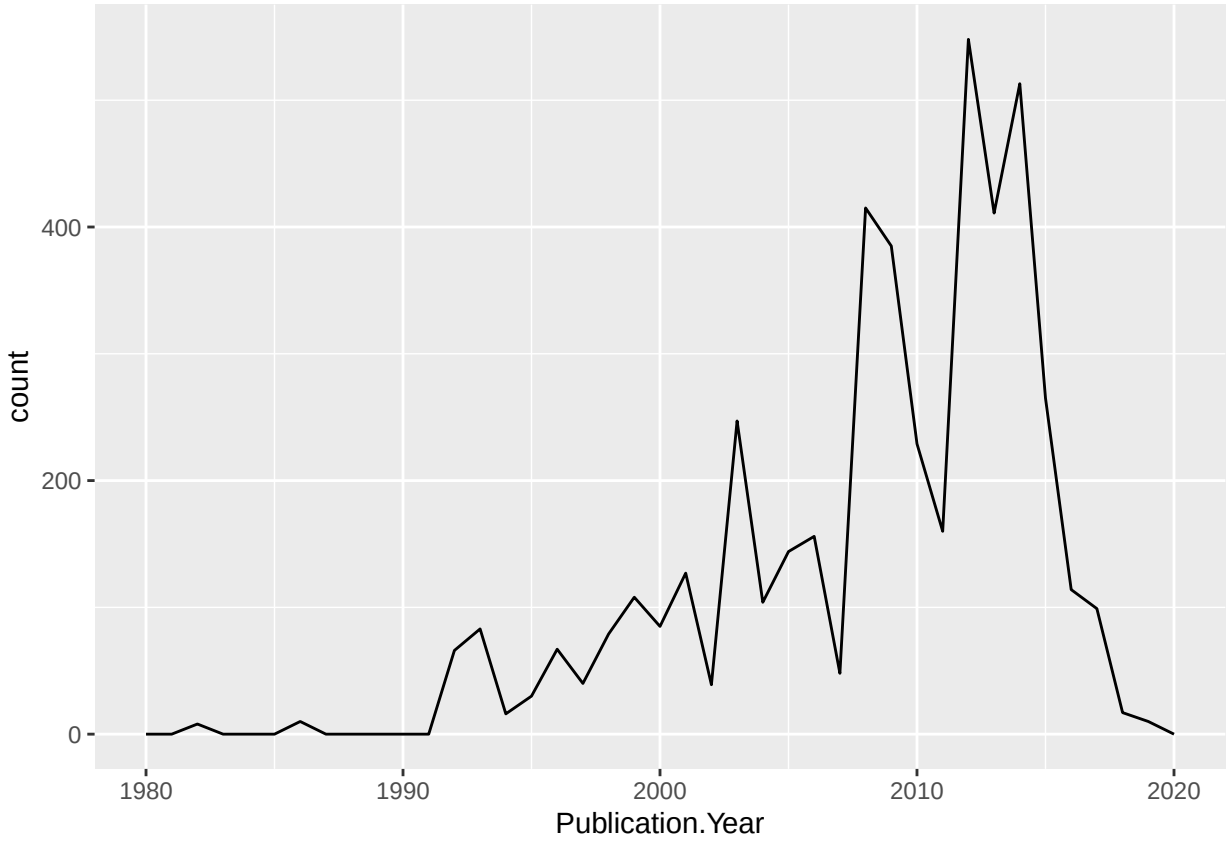
Answer: It is a char, because some of the rows have values that aren't strictly numeric, like "~10", or "12/".

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 = 41) +
  scale_x_continuous(limits = c(1980, 2020)) +
  theme(legend.position = "top")
```

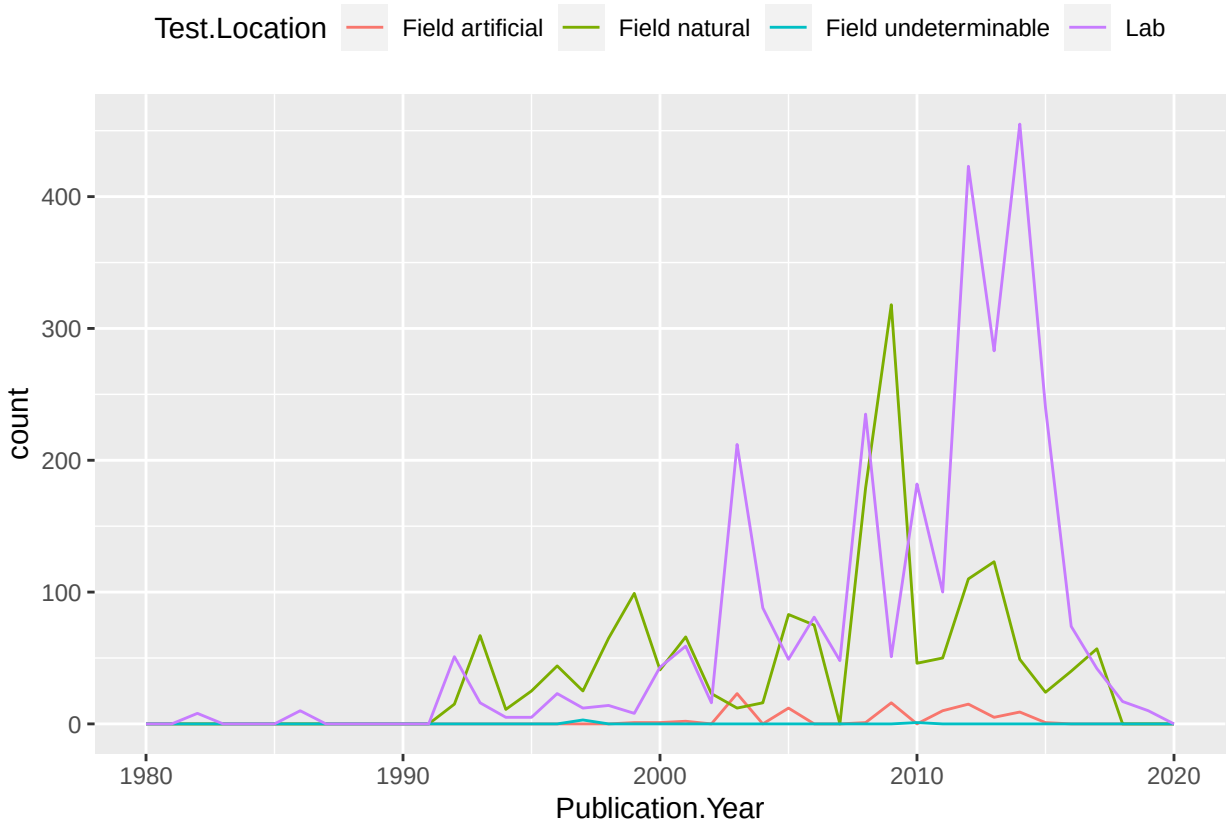
```
## Warning: Removed 2 row(s) containing missing values (geom_path).
```



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 = 41) +
  scale_x_continuous(limits = c(1980, 2020)) +
  theme(legend.position = "top")
```

```
## Warning: Removed 8 row(s) containing missing values (geom_path).
```

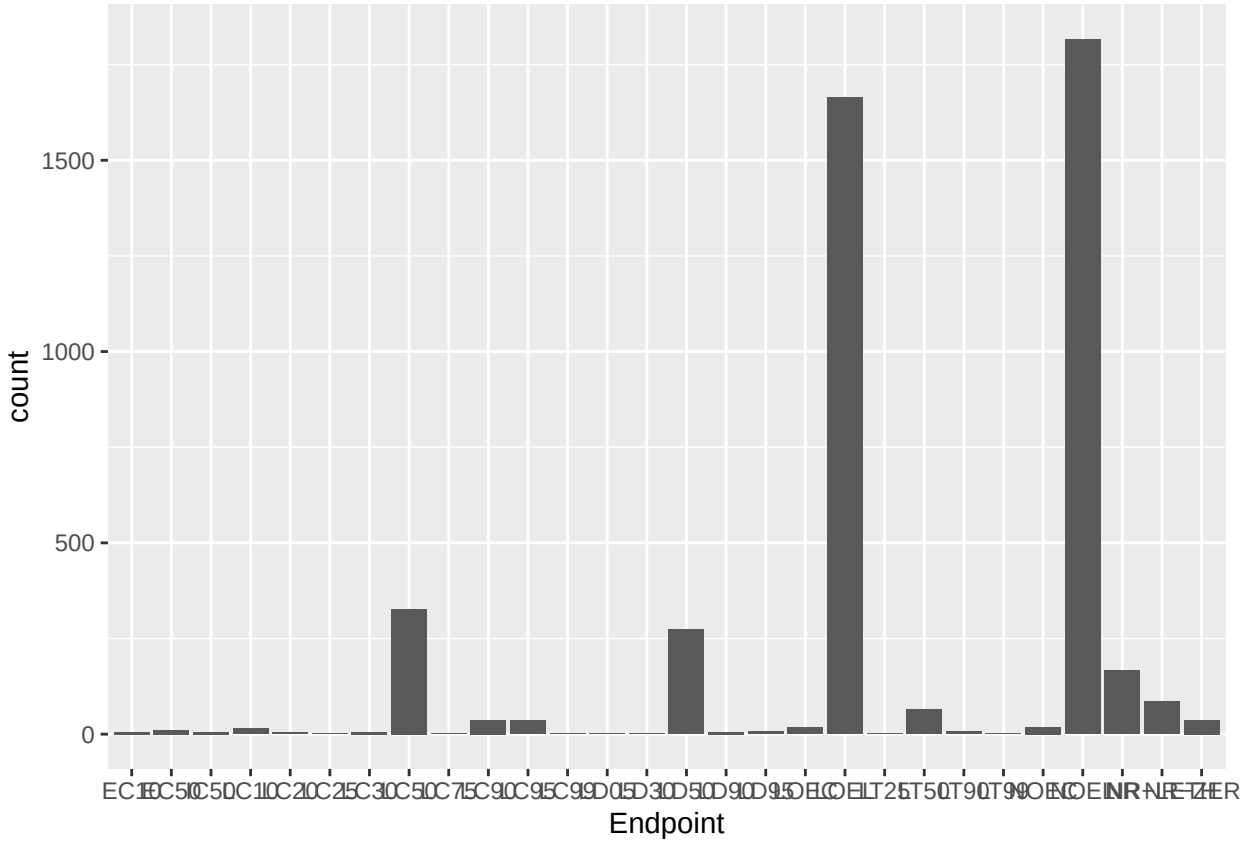


Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: Lab and Field Natural. Most years Labs are the most common locations, but for some years, natural field studies are more common.

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()
```



Answer: LOEL and NOEL, LOEL being "Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls", and NOEL being "No-observable-effect-level: highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test". Note that the x-axis labels cannot be read when printed on a pdf, and can only be seen when the graph is zoomed in on in R.

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.

```
#summary(Litter$collectDate)
#based on this, it's a char.
#using lubridate to update it to a date format
library(lubridate)
```

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

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

```
Litter$collectDate <- ymd(Litter$collectDate)
summary(Litter$collectDate)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
```

```
## "2018-08-02" "2018-08-02" "2018-08-30" "2018-08-16" "2018-08-30" "2018-08-30"
```

```
unique(Litter$collectDate)
```

```
## [1] "2018-08-02" "2018-08-30"
```

```
#two days, 08-02 and 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`?

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

```
## [1] 12
```

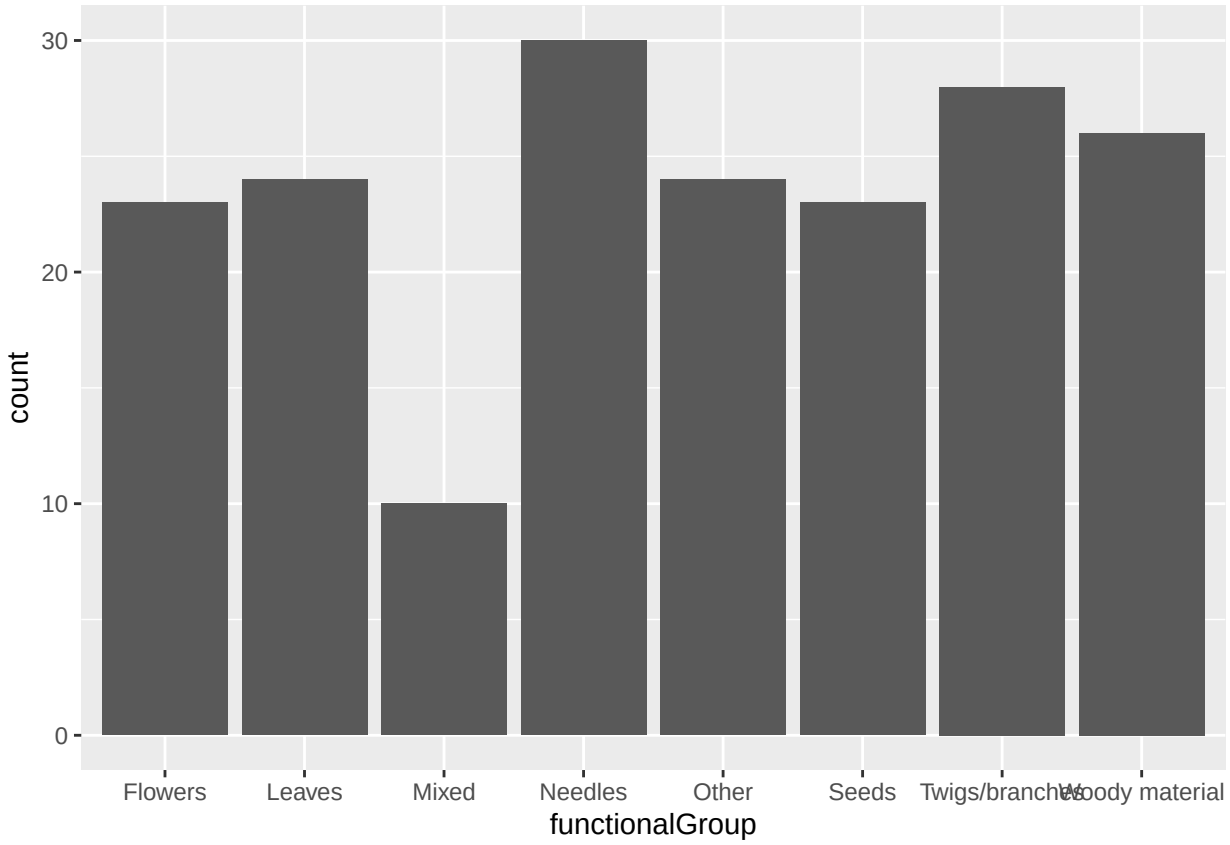
```
summary(Litter$plotID)
```

```
##    Length     Class      Mode
##       188 character character
```

Answer: 12 plots. summary gives the number of rows, but does not evaluate how many are repeats or unique.

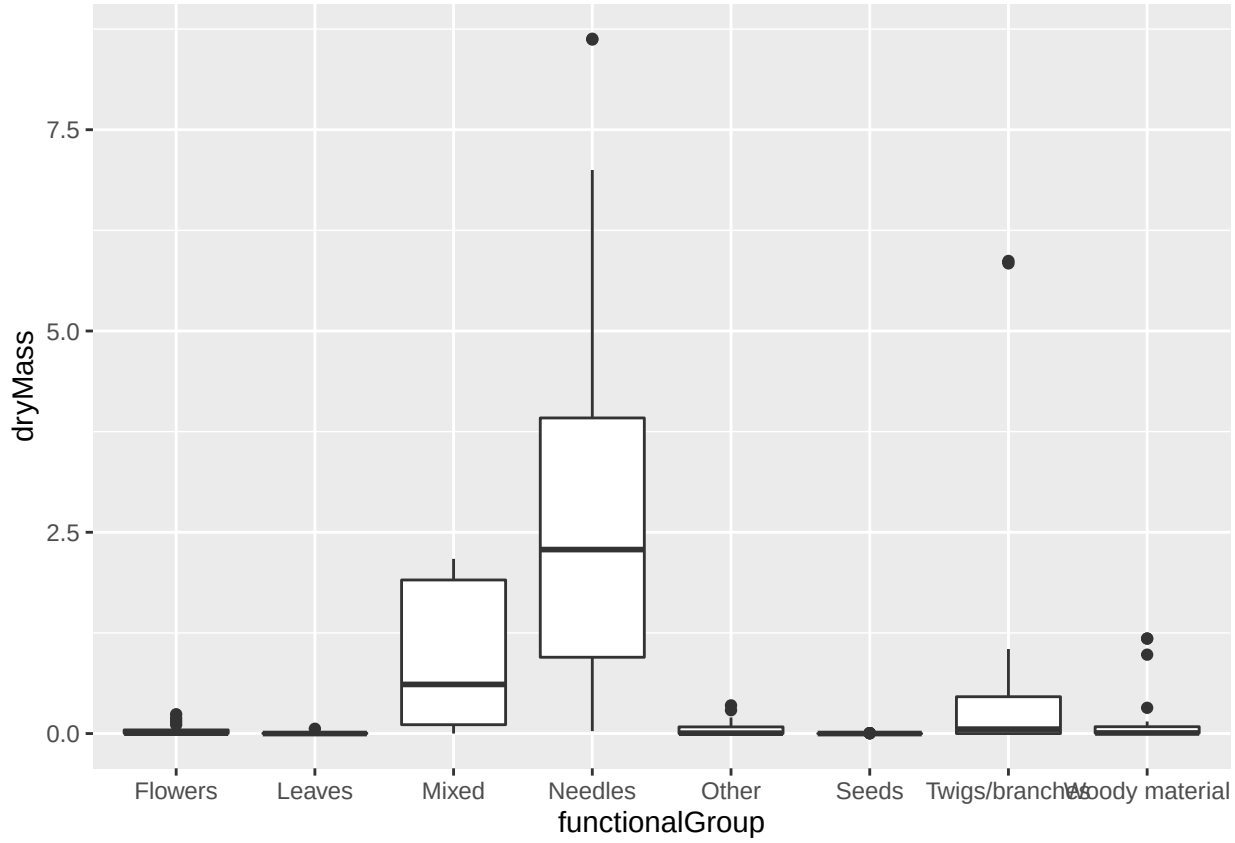
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
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))
```

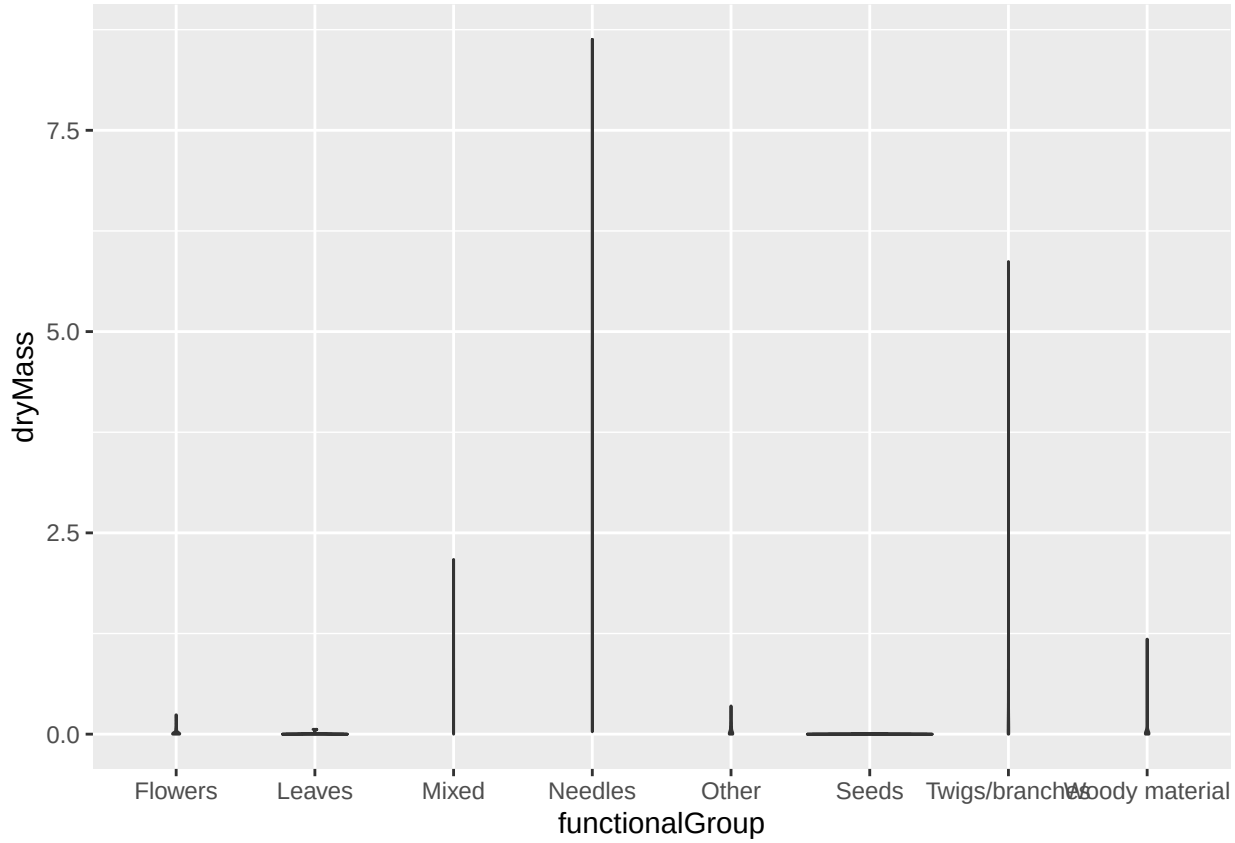



```
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75))
```

```
## 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: There isn't a sufficient density of repeat values to make the violin plots get "fat", so it's just a line.

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

> Answer: Needles