Assignment 3: Data Exploration

Tommy Hancock

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 Tuesday, January 28 at 1:00 pm.

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.

```r
getwd()
```

```
## [1] "C:/Users/thoma/Thomas/2018 Grad School/Duke MEM/ENV 872/Environmental_Data_Analytics_2020/Assign
```

```r
library(tidyverse)
Neonics <- read.csv("../Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv")
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: Since many insects serve as pollinators, widespread use of pesticides (such as neonicotinoids) could impact food production by decreasing the number of pollinators available to pollinate plants on farms. This phenomenon is already being observed with bees. There are also ecological preservation concerns since many animals eat 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: Leaf litter provides a lot of the nutrients for the forest soil, and therefore can serve as an indication of how healthy the soil (and the entire forest ecosystem) is. Leaf litter can also provide indications of the carbon flux in a forest over time (e.g., is the forest storing lots of carbon).

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: The litter and woody debris are collected from traps (elevated for litter and ground-level for woody debris). For each trap, the material is sorted into 8 catagories, and each category is weighed and reported. Each elevated (litter) trap is 0.5 square meters in size. The ground traps (woody debris) are 3m x 0.5m. Traps are placed to ensure they are under proper vegetation. Ground traps are sampled once per year, elevated traps are sampled throughout the year, depending on vegetation type. * Weights are reported with an accuracy 0.01 g. Detected categories with a weight <0.01g are reported as "<0.01g" indicating the detection limit. * Sampling plots are either 40m x 40m or 20m x 20m. There is one litter trap pair for every 400 square-meter plot. *Trap placement is sometimes random, sometimes targeted (depending on vegetation) to ensure the traps are below appropriate vegetation.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

   Answer: There are 4623 entries (rows) with 30 datafields (columns) per entry

6. Using the `summary` function, 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: The most common effects studied were mortality and population. These effects are likely popular for studying because they demonstrate the toxicity of the chemical on the insect (i.e., does it kill the insect). These effects are also relatively easy to measure (did the insect die, how many insects are left out of the original sample population, etc.).

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, maxsum = 20)
```

```
##                    Honey Bee                Parasitic Wasp 
##                          667                           285
##       Buff Tailed Bumblebee         Carniolan Honey Bee
##                         183                         152
##                 Bumble Bee           Italian Honeybee
##                         140                         113
##             Japanese Beetle           Asian Lady Beetle
##                          94                          76
##              Euonymus Scale                    Wireworm
##                          75                          69
##           European Dark Bee           Minute Pirate Bug
##                          66                          62
##         Asian Citrus Psyllid               Parastic Wasp
##                          60                          58
##      Colorado Potato Beetle              Parasitoid Wasp
##                          57                          51
##          Erythrina Gall Wasp                Beetle Order
##                          49                          47
## Snout Beetle Family, Weevil                     (Other)
##                          47                        2272
```

> Answer: The six msot common species are the Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carolinian Honey Bee, Bumble Bee, and Italian Honeybee. These are all types of bee (except the parasitic wasp) and they all have use in agriculture. Since they are important for agriculture (especially pollination), there is large societal motivation to study these insects in order to ensure people have food to eat.

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"
```

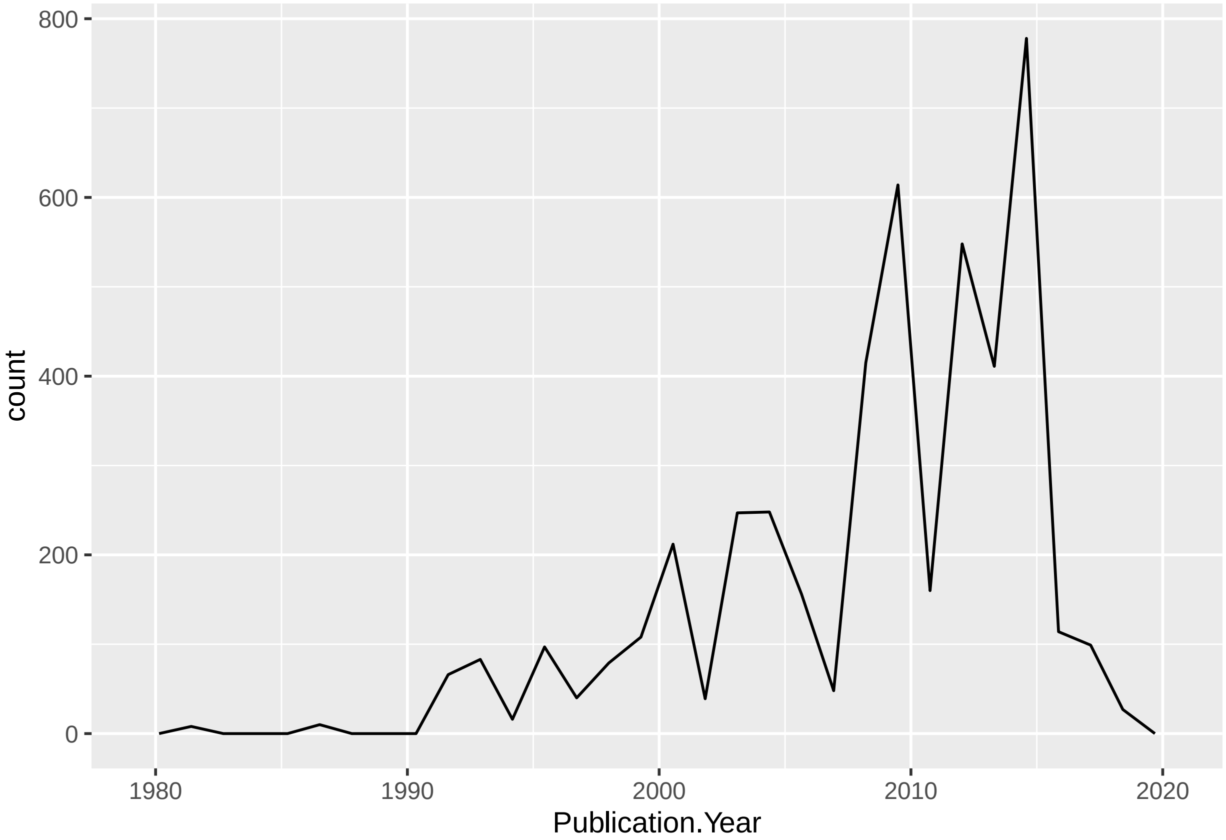
> Answer: Conc.1..Author is a factor. This is likely because there are string values (e.g., "NR") reported in the column, many of which are repeated many times. As such, these repeated values are grouped into a level or factor instead of being treated as a number or NA.

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))
```

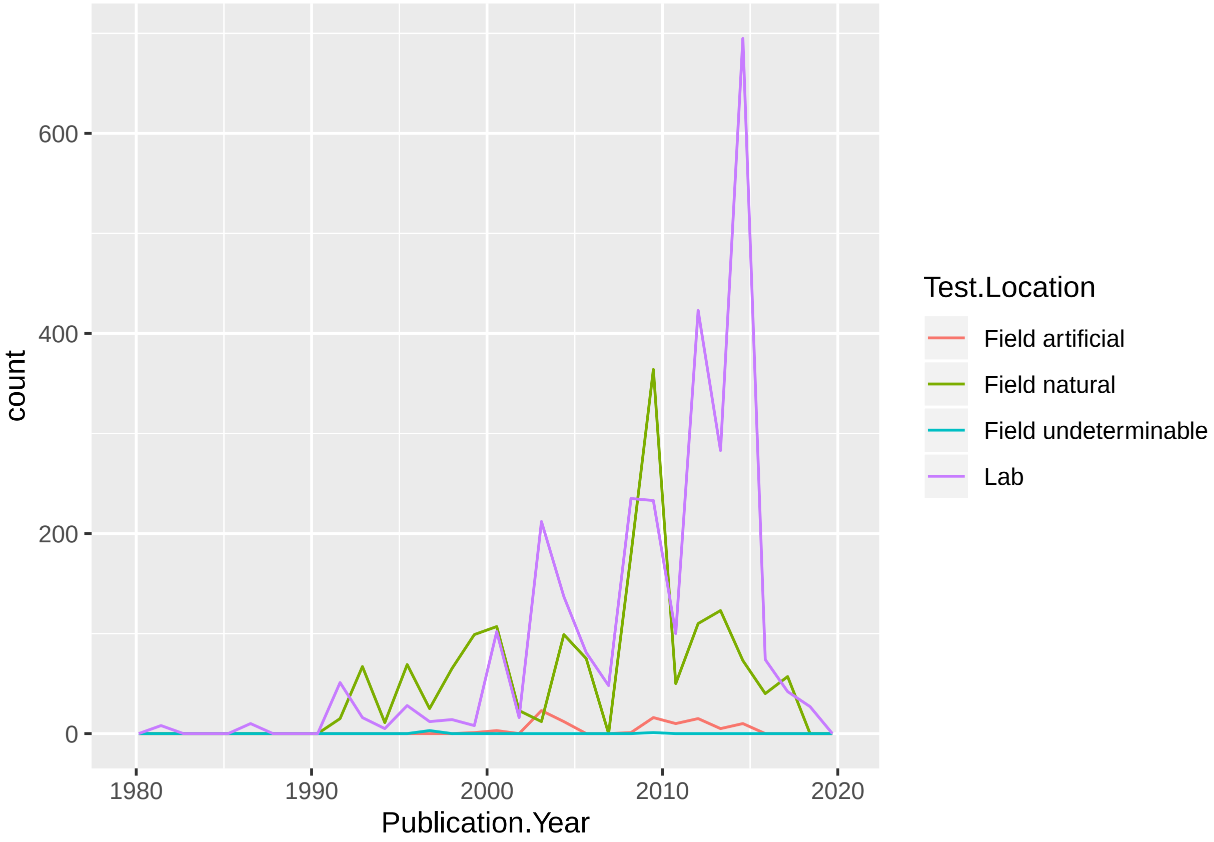
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

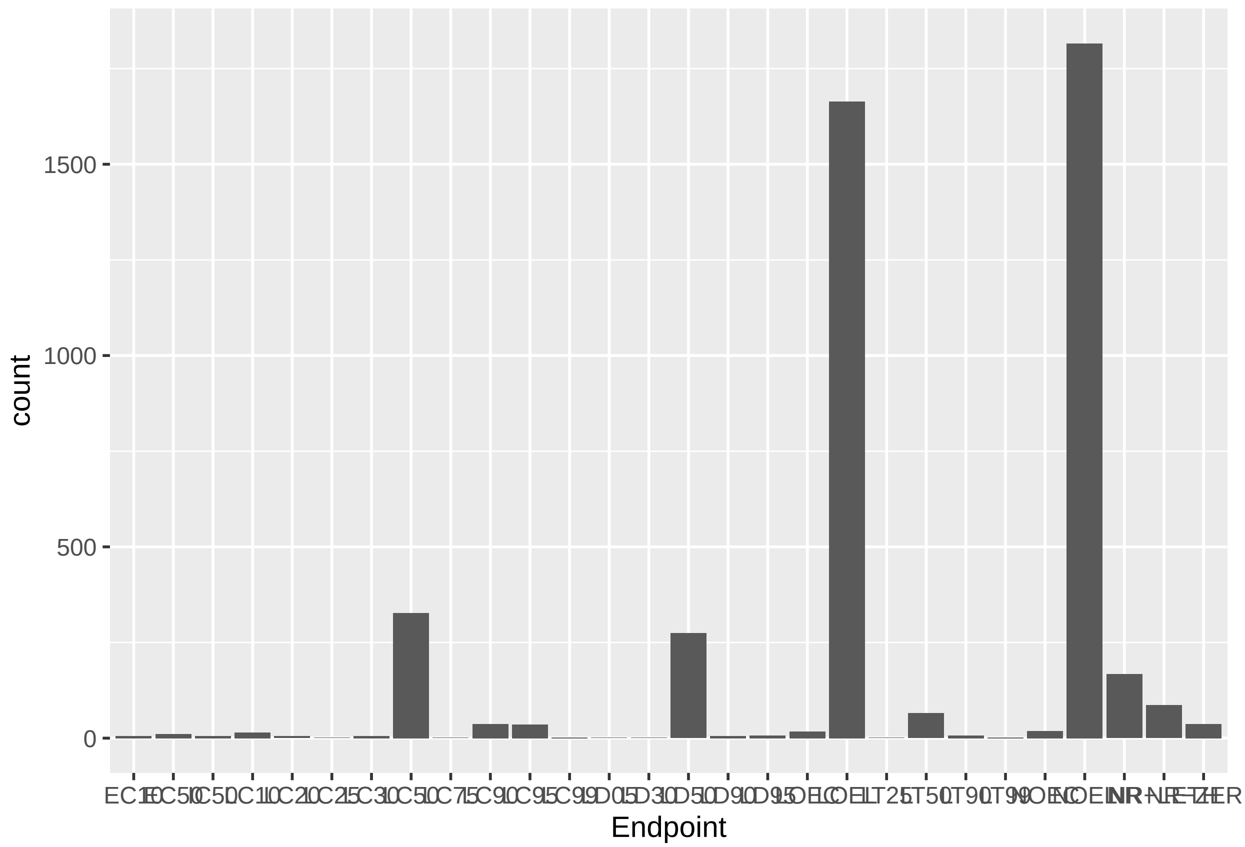


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are laboratories and natural fields. From 1990 to about 2000, natural fields were usually the most common. However, from about 2002 onwards, lab-based studies were almost always the most common (except 2009 or so).

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: The two most common endpoints are NOEL (No observable effect level) and LOEL (Lowest observable effect level). NOEL is when the highest dose does not produce effects that are significantly different from the control response. LOEL is when the lowest dose creates significantly different responses than the 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) # Check class type
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate) # Assign as "date" class
class(Litter$collectDate) # Confirm class type
```

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

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

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

Answer: There were only 2 sampling dates: 2018-08-02 and 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
```

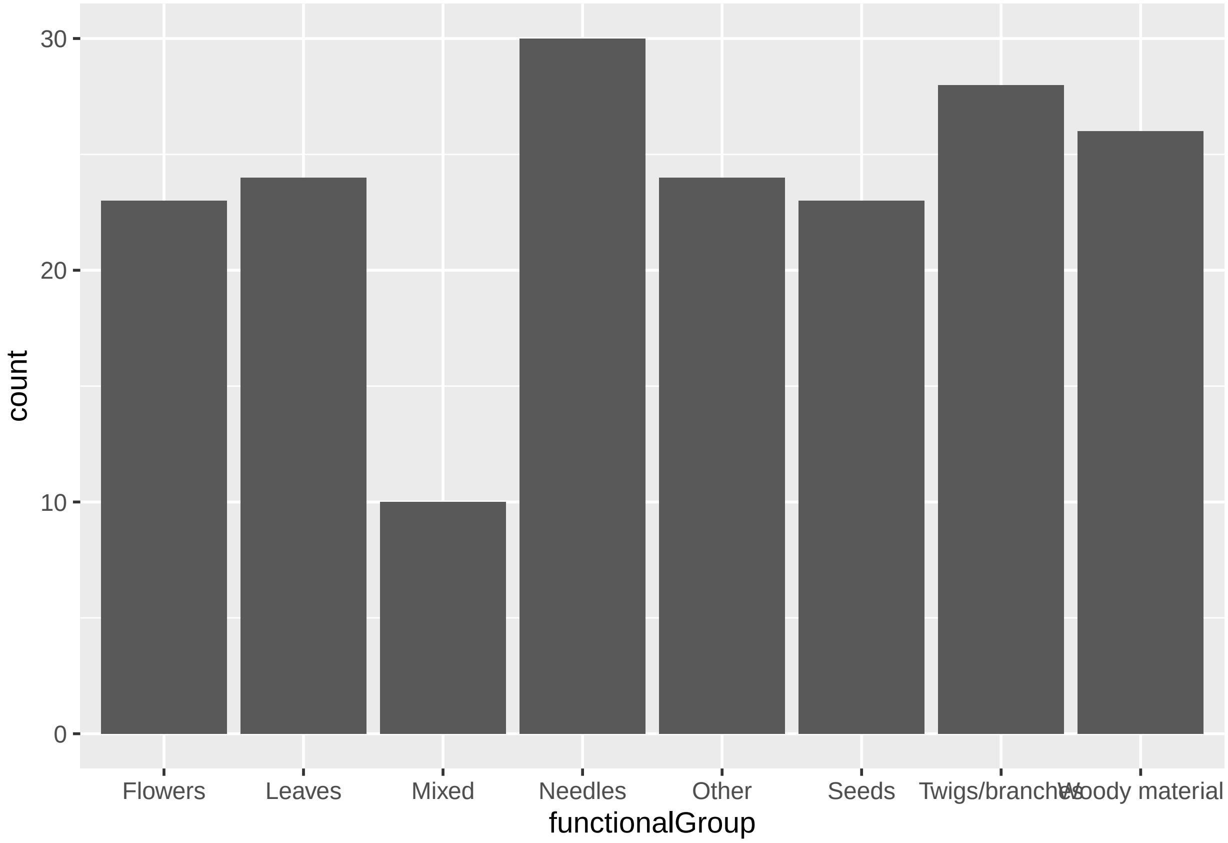
```
summary(Litter$plotID)
```

```
## 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: 12 unique plots were sampled at Niwot Ridge. “unique” provides the different levels, whereas “summary” provides the levels and the count of occurences of each level.

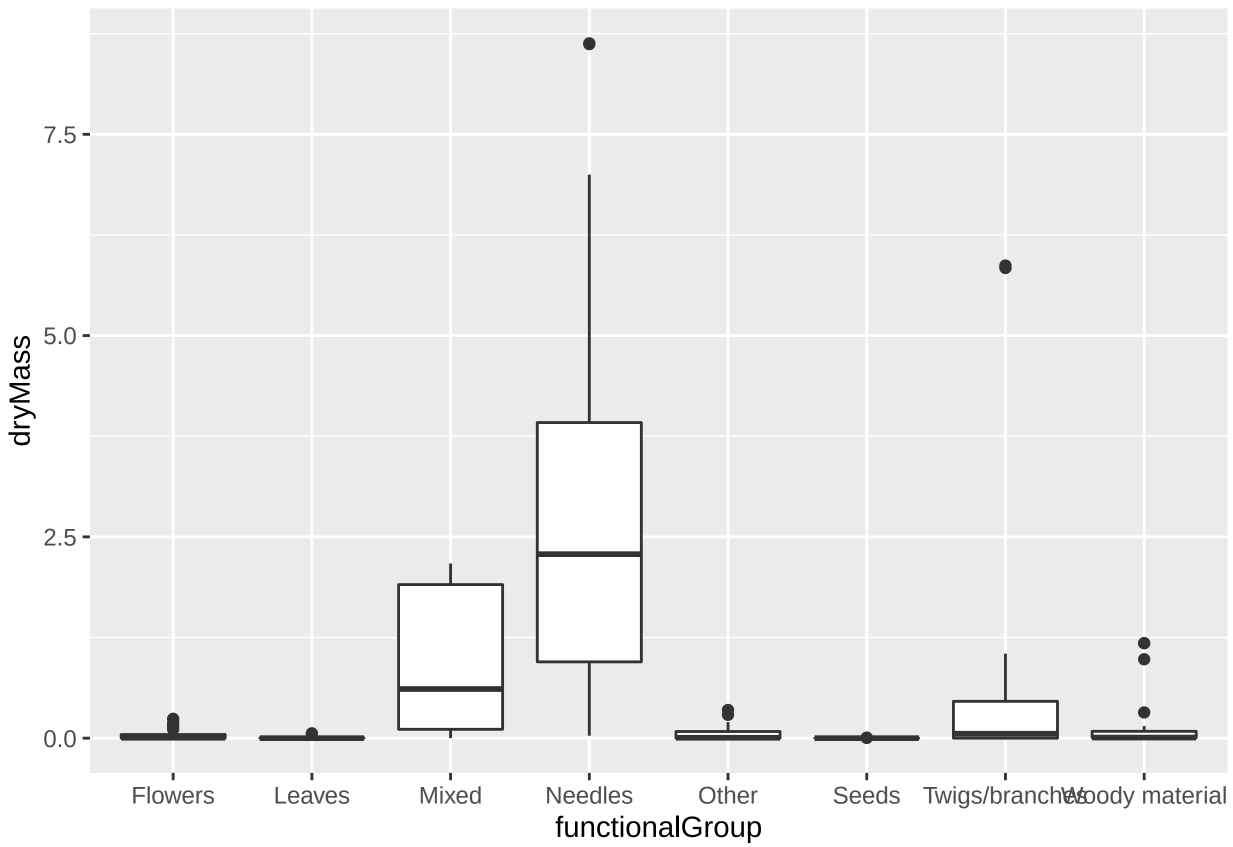
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

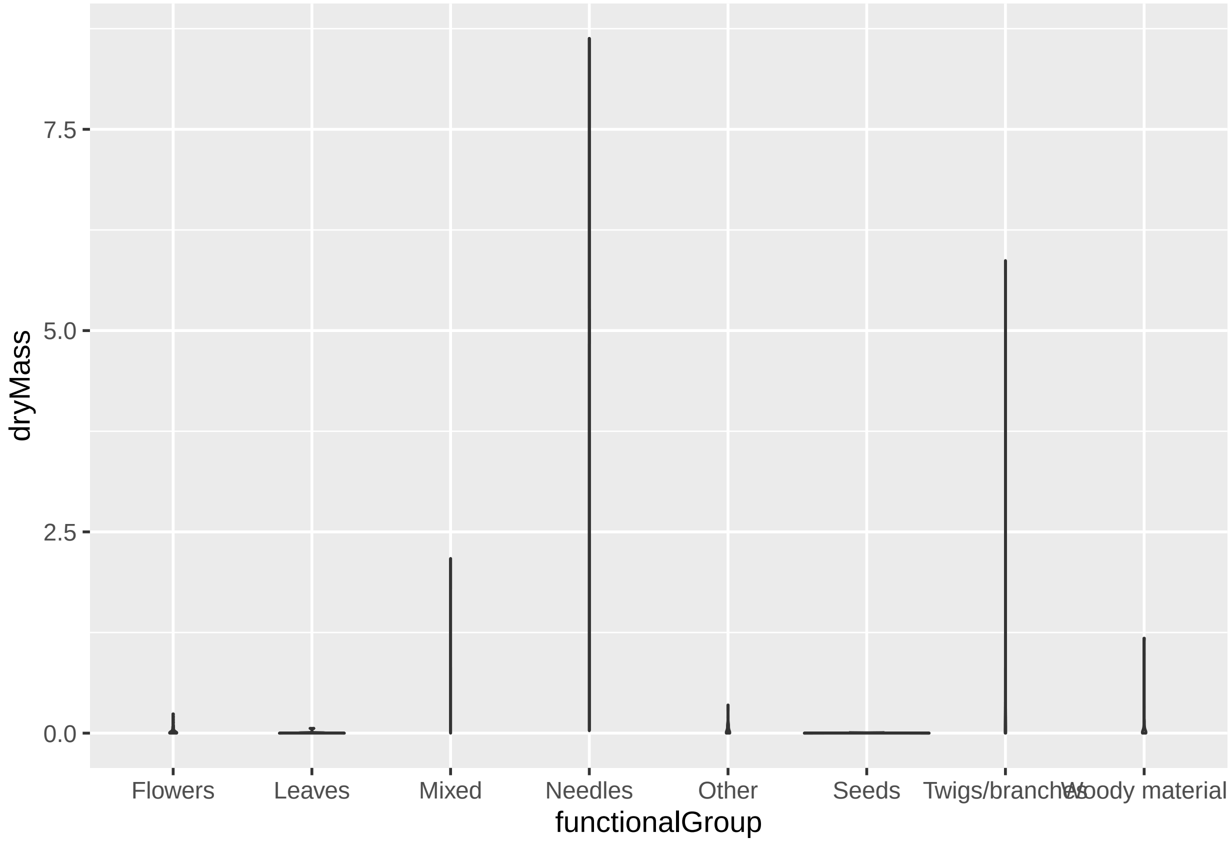


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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



```r
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: In this case, the violin plots are too narrow to really be seen (the distribution function has very long tails compared to the mass, if I understand how it works correctly). Since the boxplots are a fixed width, you can see which ones have higher and lower median masses, etc. (This could potentially be solved by scaling the width so they all have the same width.) Because the violin plots appear as more-or-less just a line, it is impossible to tell if the line is long because of outliers or if the mass is distributed along it.

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

> Answer: The "Needles" and "Mixed" types of litter tend to have the highest biomass.