Assignment 3: Data Exploration

Anne Harshbarger

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.

```
getwd()
```

```
## [1] "C:/Users/Anne/Documents/ENV872/Environmental_Data_Analytics_2021/Assignments"
```

```
#install.packages("tidyverse", "ggplot2")
library(tidyverse, ggplot2)
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: We need to understand the species-specific effects of neonicotinoids on insects that are not the target species for neonicotinoid applications in addition to the ones that are. This information can tell us whether the insecticide will be effective against the species that are considered pests to the crops in fields where it is being applied, while also allowing us to make inferences about the effects on native insects and other non-pest insects in the surrounding ecosystem. In particular, pollinator species are vulnerable to the effects of pesticides, and because they are essential to ecosystem function and the successful reproduction of both agricultural and non-agricultural plants, any negative effects of neonicotinoids on pollinators need to be well understood. By understanding these effects, managers will be able to take them into account when decisions are made about the application of neonicotinoids. Like any pesticide, neonicotinoids have the potential to be transported through the ecosystem in soil or water, and they may have effects on non-insect species as well.

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 have several ecosystem functions, including creating habitat for organisms on the forest floor, providing food for decomposers and herbivores, and contributing to nutrient cycling. As leaves and other types of litter decompose, they reintriduce organic matter to the soil, which helps plants and other organisms acquire nutrients. The type and amount of leaf little present shapes forest floor communities, especially of invertebrate species. The presence of litter and woody debris can also shape the risk and impacts of fires on forest ecosystems.

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: Within the NEON network, litter and woody debris are sampled through a series of PVC basket traps on vegetated plots. Any litter that ends up in the traps is quantified by functional groups and the dry biomass (in g) is measured. 1) Paired traps are deployed to catch litter on each plot; one trap in each pair is on the ground and sampled annually, and the other is elevated above ground level and sampled biweekly and seasonally if under deciduous vegetation or monthly/bimonthly year-round if under evergreen vegetation. 2) The locations of 40 x 40 m and 20 x 20 m plots are chosen randomly within airsheds of NEON towers but must be sufficiently distant from other plots, roads, and buildings. 3) Within a plot, the locations of traps are either randomized if the plot contains >50% cover of woody vegetation that is >2 m tall, or selected to be in areas with sufficient vegetation if the vegetation cover of the plot is <50%.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
dim(Neonics) #get dimensions; 4623 rows, 30 columns
```

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

The dataset contains 4623 rows and 30 columns

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?

```
Neonics$Effect <- as.factor(Neonics$Effect) #change Effects from character to factor
summary(Neonics$Effect, maxsum=4) #summarize 4 most common effects
```

```
## Population  Mortality   Behavior    (Other)
##       1803       1493        360        967
```

Answer: Population, mortality, and behavior are the most common effects. Changing the population of the species of interest in either a positive or negative direction, causing an unusal mortality event, or causing major shift in behavior could all have major implications for the ecosystem, depending on the species that experiences the effect and the magnitude of the effect. In the case of insects, especially pollinators, a reduction in population, an increase in mortality, or a change in behavior that causes them to leave the area could negatively impact the health of plants that are pollinated by those species and any animals that rely on them as prey.

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$Species.Common.Name <- as.factor(Neonics$Species.Common.Name) #convert species to factor
summary(Neonics$Species.Common.Name, maxsum=7) #summarize top 7 (determine top 6 species after "other"
```

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

Answer: The six most common species were honey bees, parasitic wasps, buff tailed bumblebees, carniolan honey bees, bumble bees, and Italian honeybees. All of these species are pollinators. They are essential to ecosystem health because they facilitate plant reproduction, and they are not pests to agricultural crops; in fact, they can benefit crops as well as native species.

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.) #determine class
```

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

Answer: Conc.1..Author. is a character vector. It includes characters that are not numbers or decimals, including letters, /, and ~. This suggests that the concentrations in this column were not reported in a consistent way, or that there may have been unusual formatting in the csv did not import correctly. We can see from the neighboring columns that these concentrations differ in type and unit, which could lead to either cause mentioned above.

Explore your data graphically (Neonics)

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

```
ggplot(data=Neonics, aes(x=Publication.Year)) +
  geom_freqpoly(binwidth=1) +
  labs(x="Publication Year", y="Count of tests")
```

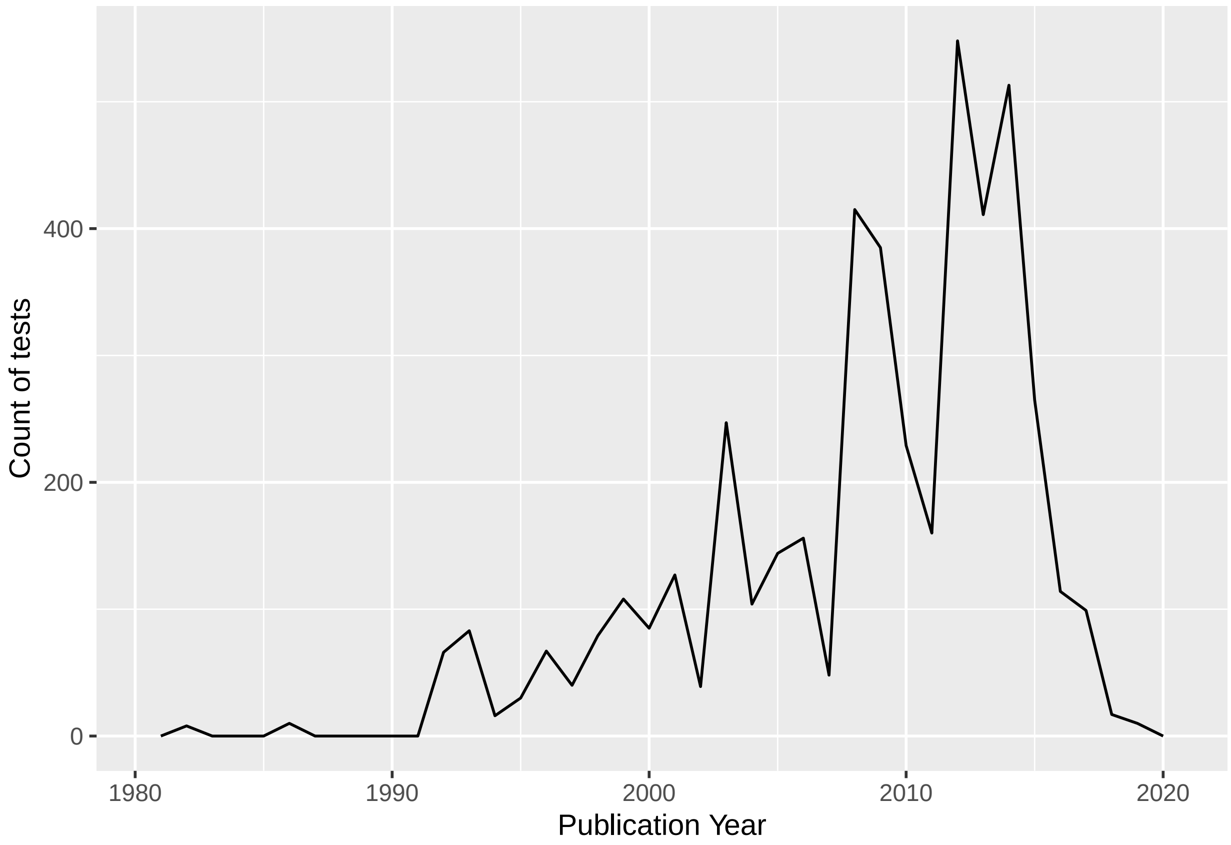

Figure 1: Nnumber of studies done on neonicotinoids and insects in each publication year.

10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
ggplot(data=Neonics, aes(x=Publication.Year, color=Test.Location)) +
  geom_freqpoly(binwidth=1) +
  theme(legend.position = "top") +
  labs(x="Publication Year", y="Count of tests")
```

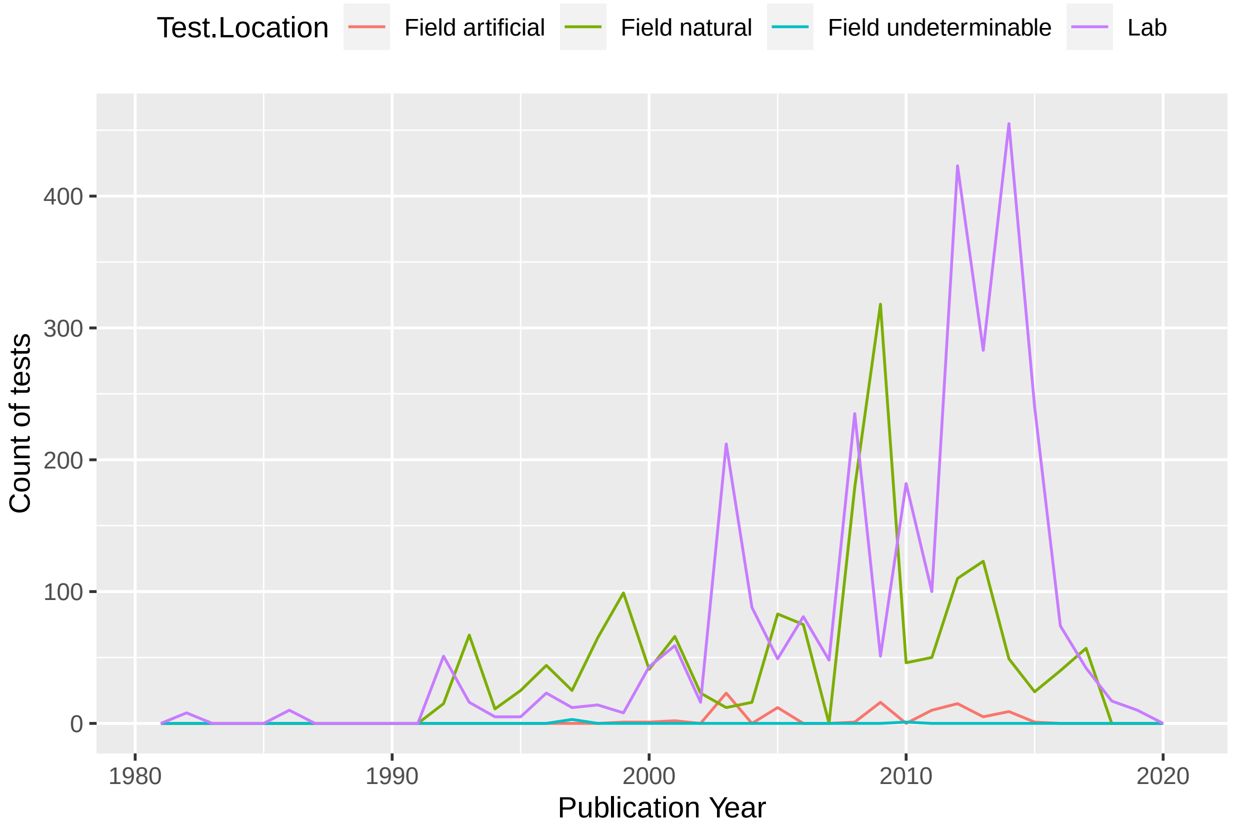

Figure 2: Number of studies done on neonicotinoids and insects in each testing location during each publication year.

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

> Answer: The most common test locations are natural field locations and lab experiments. The number of natural field tests increased slowly between 1990 and approximately 2006, then rapidly from 2006 to 2010. Around 2010, there was a major shift from natural field tests to lab tests, which continued to increase rapidly until about 2015. Artifical field tests and undeterminable field tests remained low throughout the study.

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(data=Neonics, aes(y=Endpoint)) +
  geom_bar() +
  labs(x="Number of occurrences", y="Endpoint type")
```

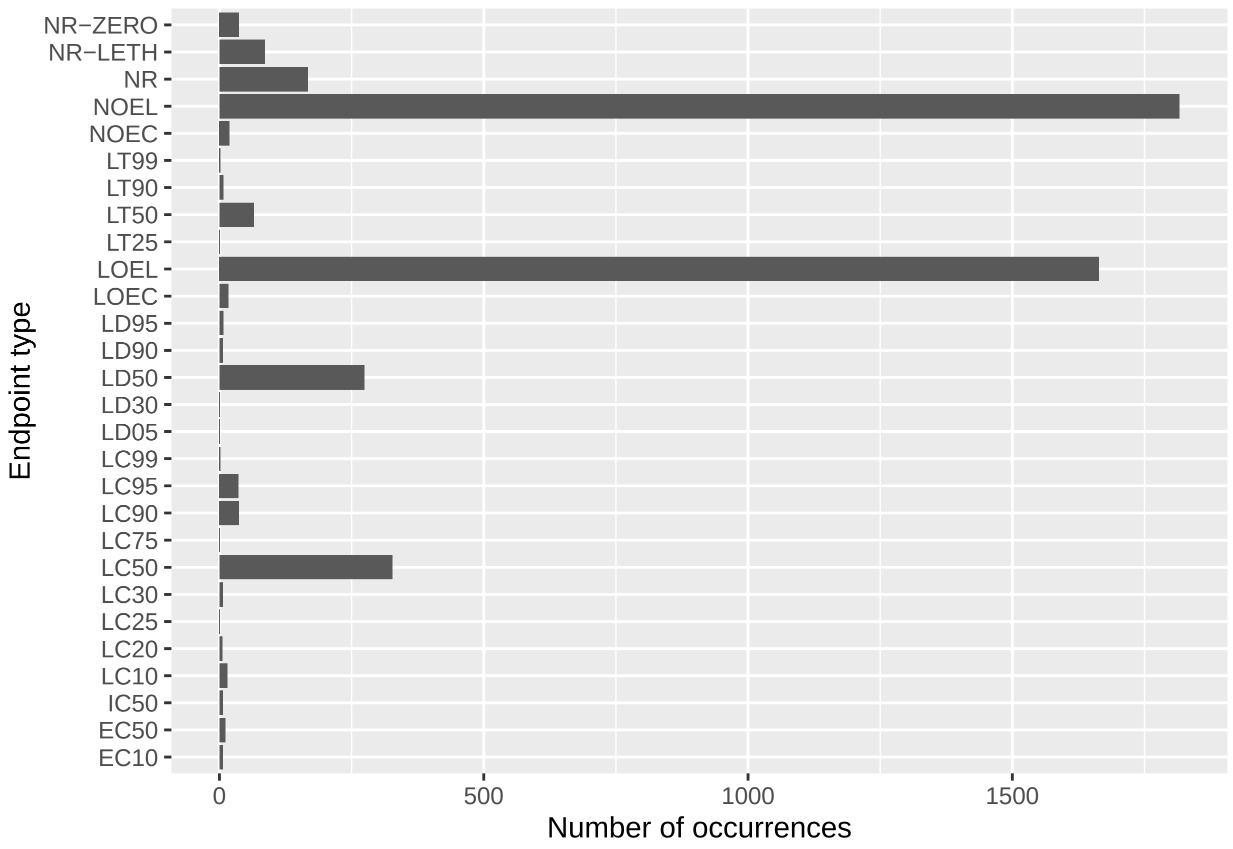

Figure 3: Number of neonicotinoid studies reaching each type of endpoint.

> Answer: NOEL (no observable effect level) and LOEL (lowest observable effect level) are the two most common endpoints. No observable effect level means that at the highest dose of treatment, the response of the organism was not significantly different from the control group. Lowest observable effect level means the lowest dose of treatment that produced a response in the treatment group that was significantly different from the control group.

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) #determine class; character
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d") #change class to date
#with format YYYY-MM-DD
class(Litter$collectDate) #determine class; date
```

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

```
unique(Litter$collectDate) #list unique collection dates
```

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

Litter was sampled on August 2, 2018 and August 30, 2018.

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) #list unique sites (12)
```

```
##  [1] "NIWO_061" "NIWO_064" "NIWO_067" "NIWO_040" "NIWO_041" "NIWO_063"
##  [7] "NIWO_047" "NIWO_051" "NIWO_058" "NIWO_046" "NIWO_062" "NIWO_057"
```

```
summary(Litter$plotID) #summarize plotID; character
```

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

```
summary(as.factor(Litter$plotID)) #summarize plotID; factor
```

```
## 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 function 'unique' gives a list of the unique elements in the vector; in this case, it returns 12 unique plot ID numbers, as 12 sites were sampled at Niwot Ridge. Because the plotID vector was imported as a character vector when the dataframe was created from the original csv, the 'summary' function initally only gives the length, class, and mode of the vector; however, after converting plotID to a factor, 'summary' gives the number of times each unique plot was sampled.

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() +
  labs(x="Functional group of litter", y="Count of sites where collected")
```

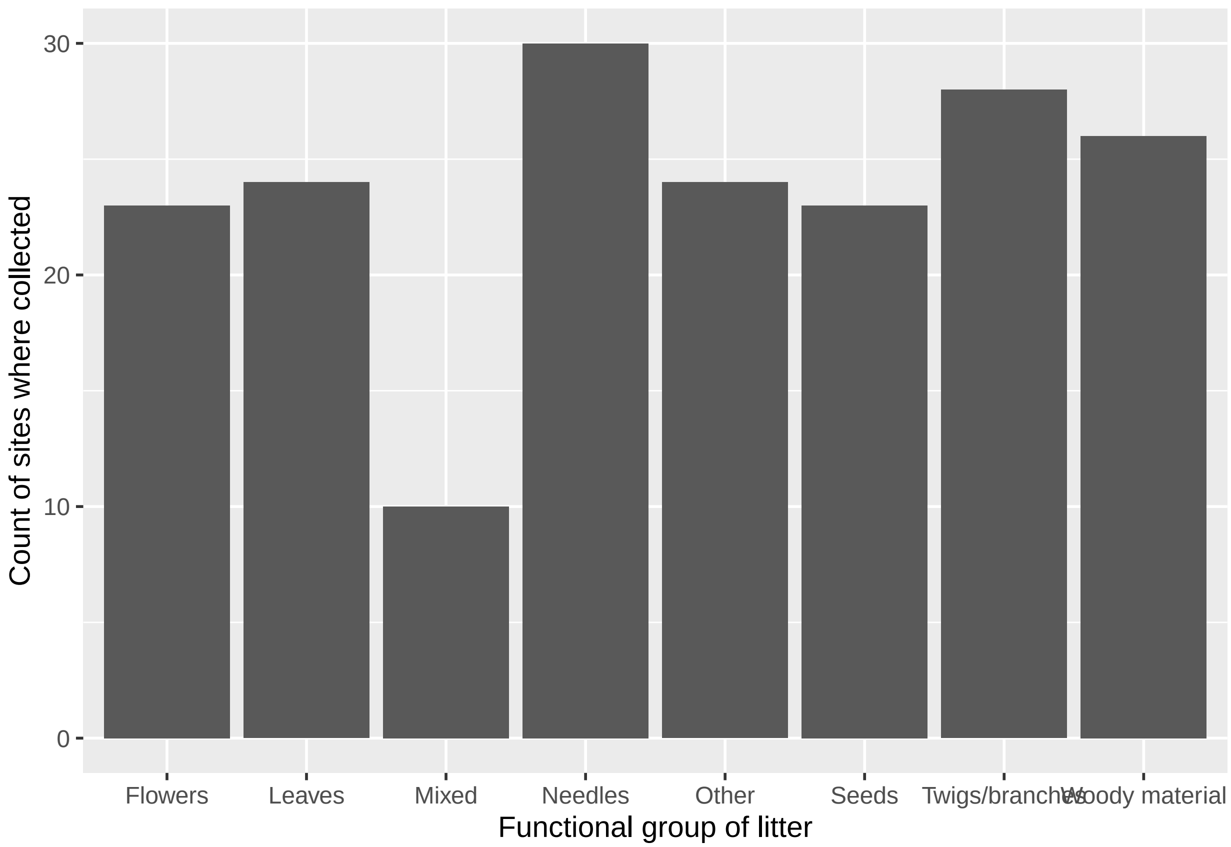

Figure 4: Types of litter, organized by functional group, found at each site on Niwot Ridge.

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

```
ggplot(data=Litter, aes(x=dryMass, y=functionalGroup)) +
  geom_boxplot() +
  labs(x="Dry mass (g)", y="Functional group of litter")
```

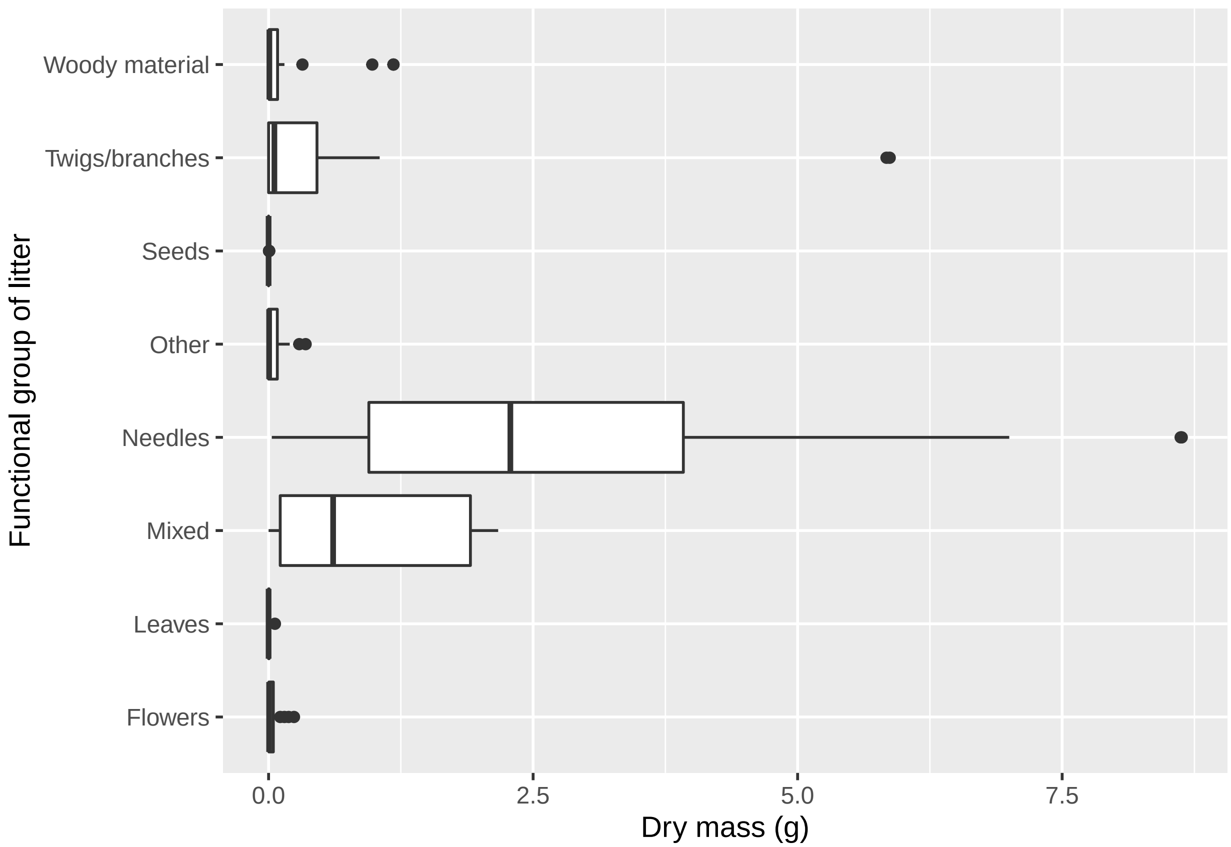

Figure 5: Boxplot of distribution of dry mass measured for each functional group.

```
ggplot(data=Litter, aes(x=dryMass, y=functionalGroup)) +
  geom_violin() +
  labs(x="Dry mass (g)", y="Functional group of litter")
```

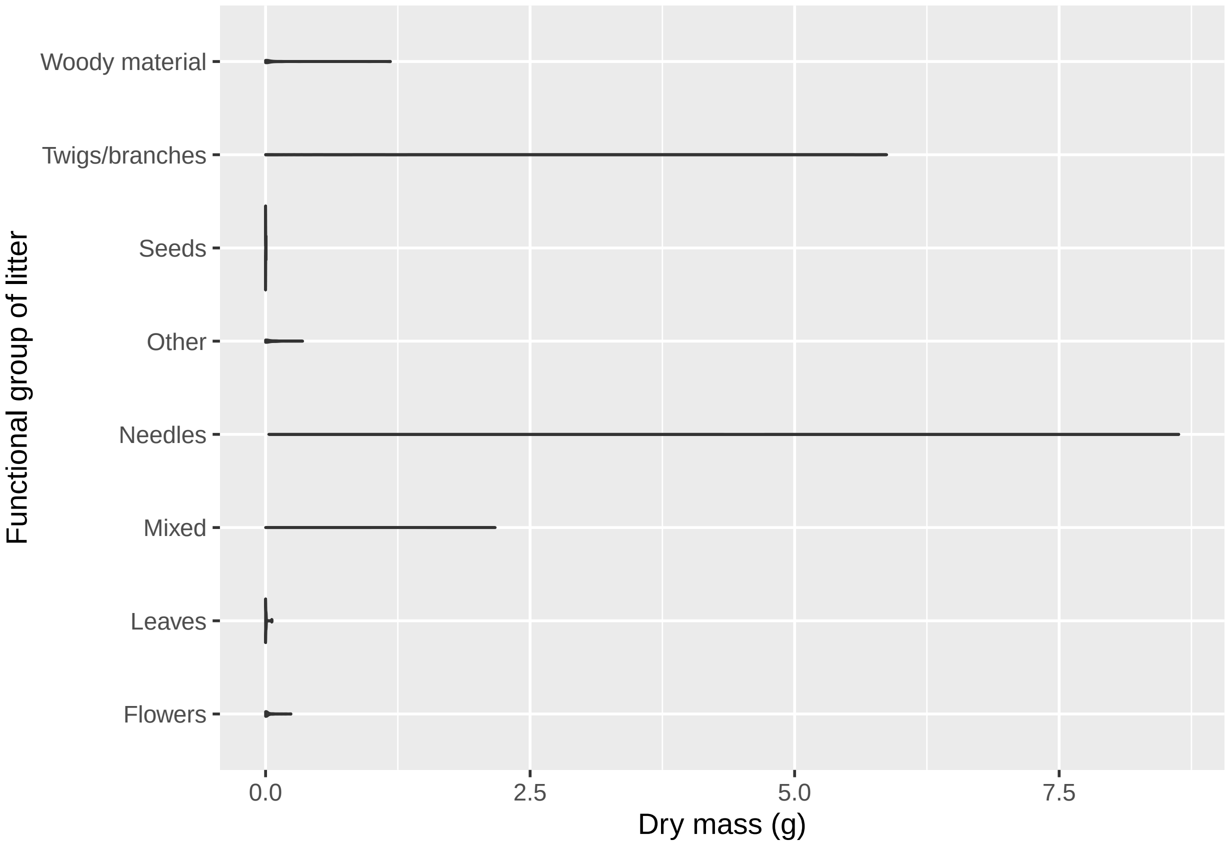

Figure 6: Violin plot of distribution of dry mass measured for each functional group.

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

> Answer: Because the distribution of values are so spread out, the violin plot displays as mostly horizontal lines (with the notable exceptions of seeds and leaves, for which points are closely clustered and the violin plots appear to be vertical lines). Given the scale and the distribution of the data, the boxplot is more readable; being able to identify the quartiles gives a better understanding of the dry mass data. Additionally, the boxplot shows outliers, while the violin plot makes it difficult to identify gaps between the outliers and where most of the datapoints occur. Note for both figures: orientation (functional group on the y-axis and dry mass on the x-axis) was chosen to improve readability of labels, boxes, and violins.

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

> Answer: Needles have the highest biomass at the Niwot Ridge sites, followed by mixed litter. This suggests the presence of many coniferous trees at these sites.

Duke Community Standard affirmation: I have adhered to the Duke Community Standard in completing this assignment. -Anne Harshbarger