

Assignment 3: Data Exploration

Amanda Booth, Section #4

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. 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] "/Users/amandabooth/Documents/GitHub/Environmental_Data_Analytics_2022/Assignments"

# setwd("/Users/amandabooth/Documents/GitHub/Environmental_Data_Analytics_2022")

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 ecotoxicology of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

Answer: Insects can pass on neonicotinoids to untreated plants and disrupt our food webs. Additionally, when pollinators are hurt by neonicotinoids, that can also disrupt our food webs.

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 impact the nutrient cycle of soil.

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: * Sampling happens at terrestrial NEON sites with woody vegetation over 2 meters tall.
* Within these sites, tower plots to sample are selected at random. * Ground traps are sampled once a year, whereas elevated traps are sampled every two weeks in deciduous forests and every one to two months at evergreen sites.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

Answer: 4623 observations of 30 variables

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: The most common effects studied are population (1803) and mortality (1493). These effects are likely of interest because they ultimately determine the damage neonicotinoids can do to species.

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: The 6 most common species studied are the honey bee (667), parasitic wasp(285), tailed bumblebee (183), carniolan honey bee (152), and italian honeybee (113). All of these species are pollinators, which may be more interesting to study over other insects because they're useful to many plants.

Note: I had to use '7' instead of '6' in my code to get the top 6 species, because the last species pulled up is 'Other'.

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"
```

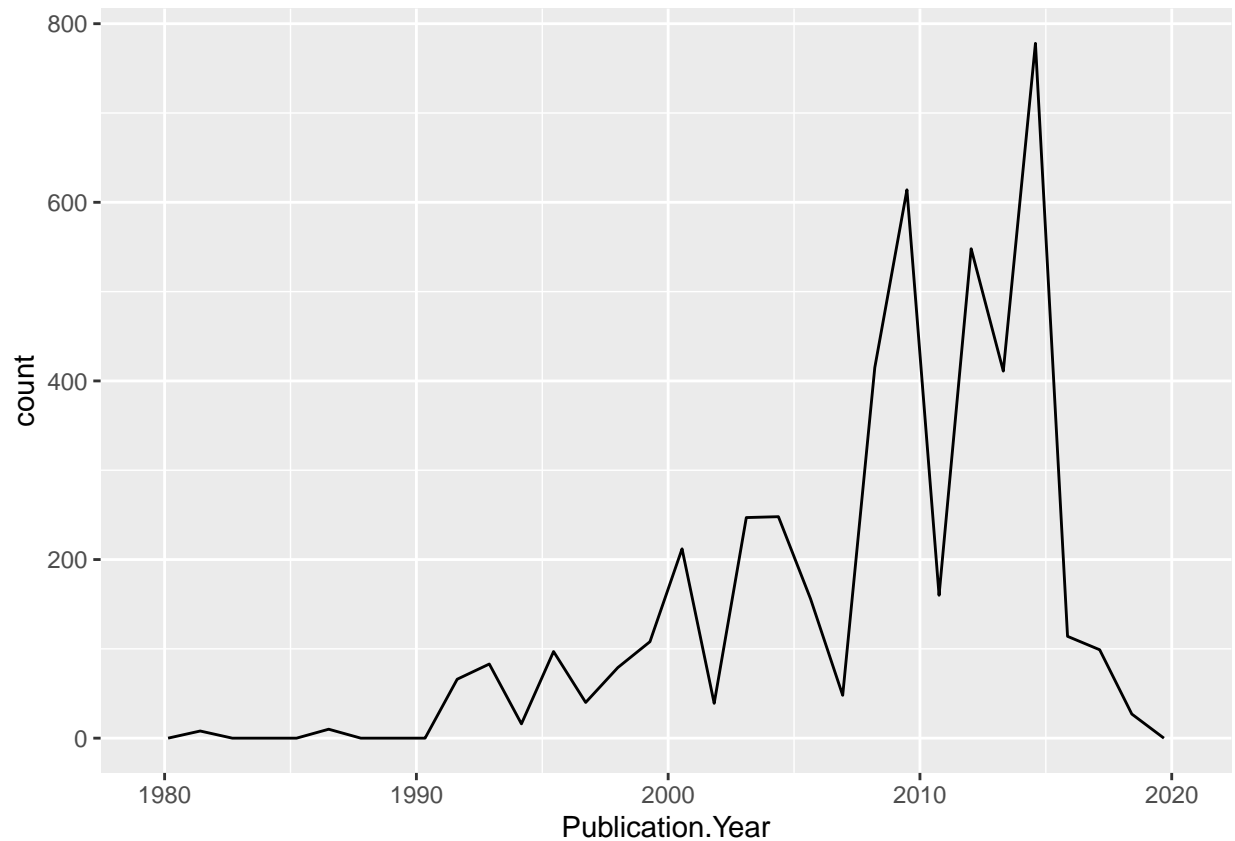
Answer: 'Conc.1..Author' has a factor (and not numeric) class because it contains characters.

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))
```

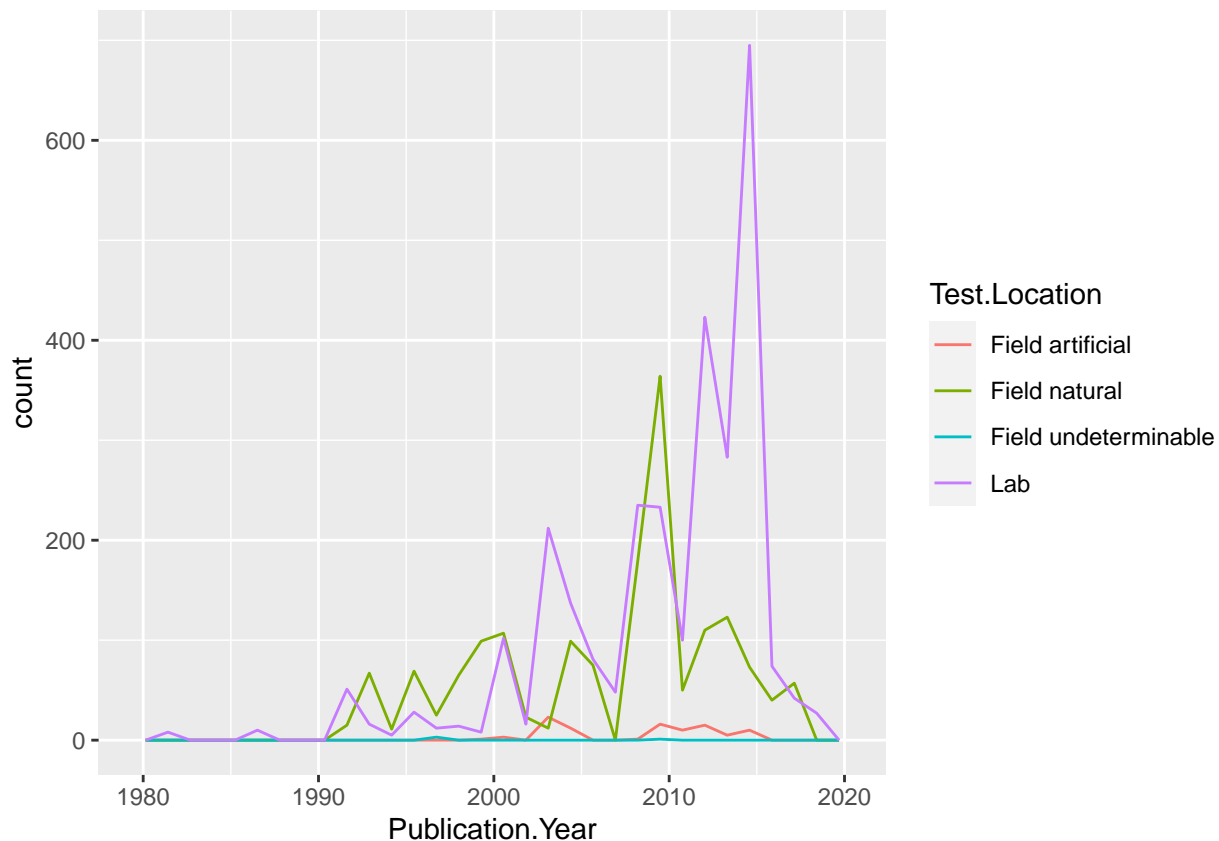
```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



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'.
```

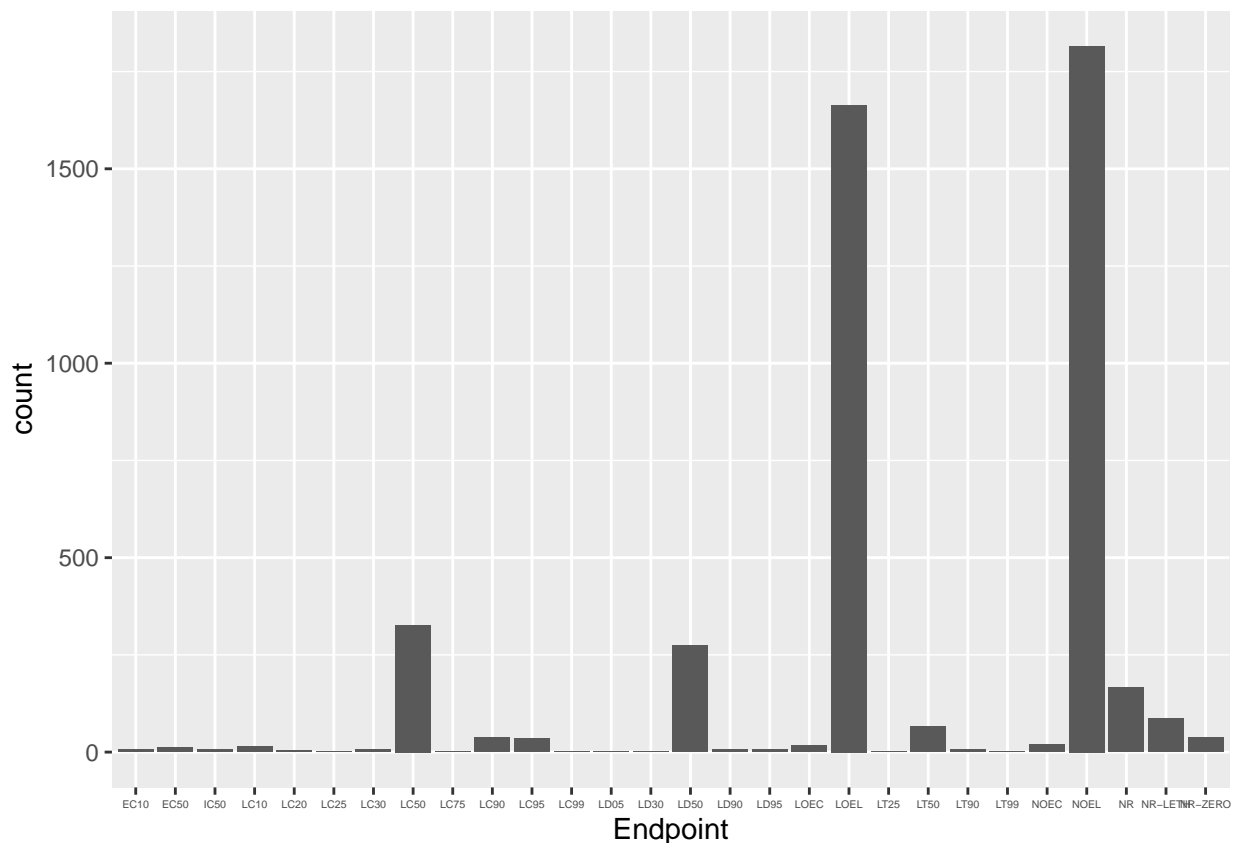


Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: Labs and natural fields are the most common test locations. While from 1990 to around 2010 they went back and forth as the most common, lab tests have shot up in popularity since then. Both seem in decline in the late 2010s, likely because studies from that time haven't been published yet.

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 = 4))
```



Answer: LOEL is defined as “lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC)” and NOEL is defined as “No-observable-effect-level: highest dose (concentration) producing effects not significantly different from responses of controls according to author’s reported statistical test (NOEAL/NOEC)”

Explore your data (Litter)

- 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)
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate, "Y%-m%-d%")
```

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

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

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

```
## [1] NA
```

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
```

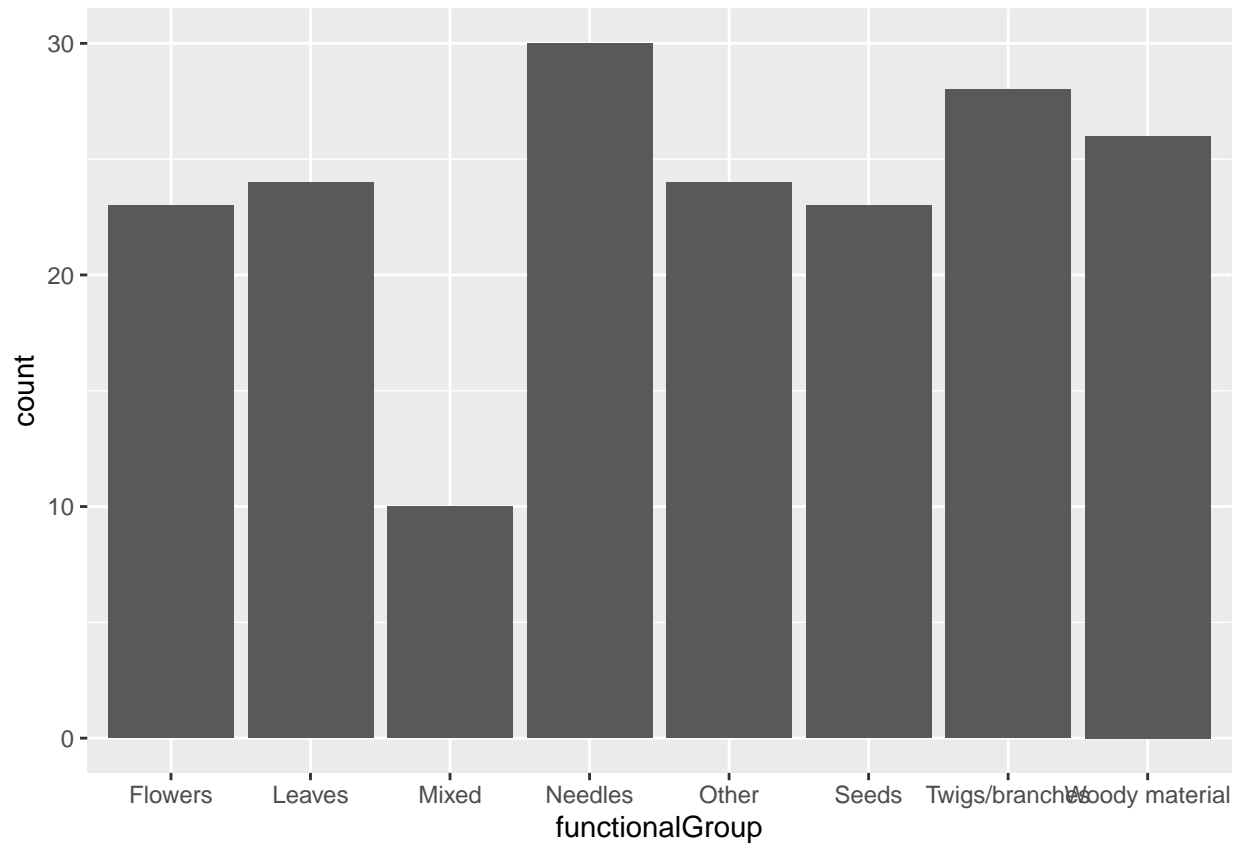
```
summary(Litter$plotID)
```

```
## 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: The ‘`unique`’ function displays the unique IDs and how many unique IDs there are. In contrast, the ‘`summary`’ function displays the unique IDs and the count for each unique ID.

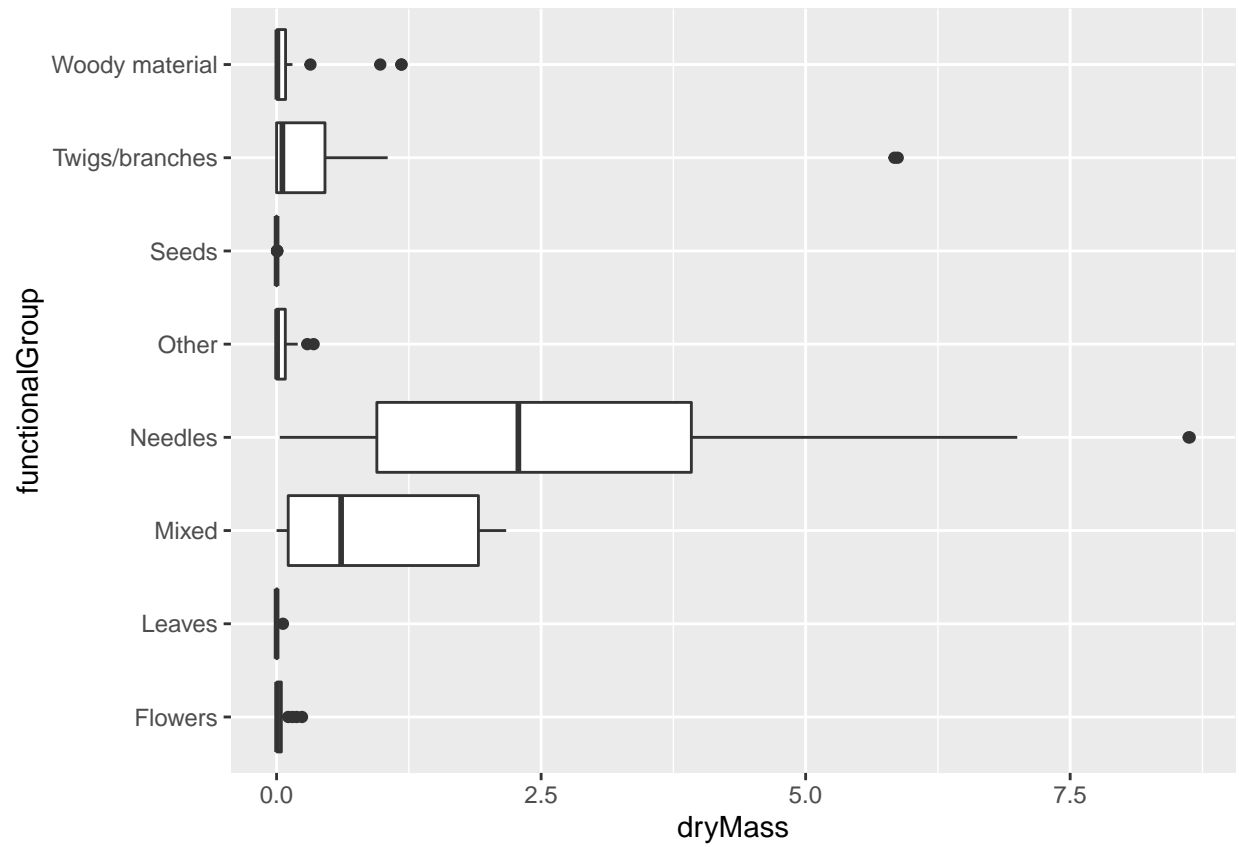
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

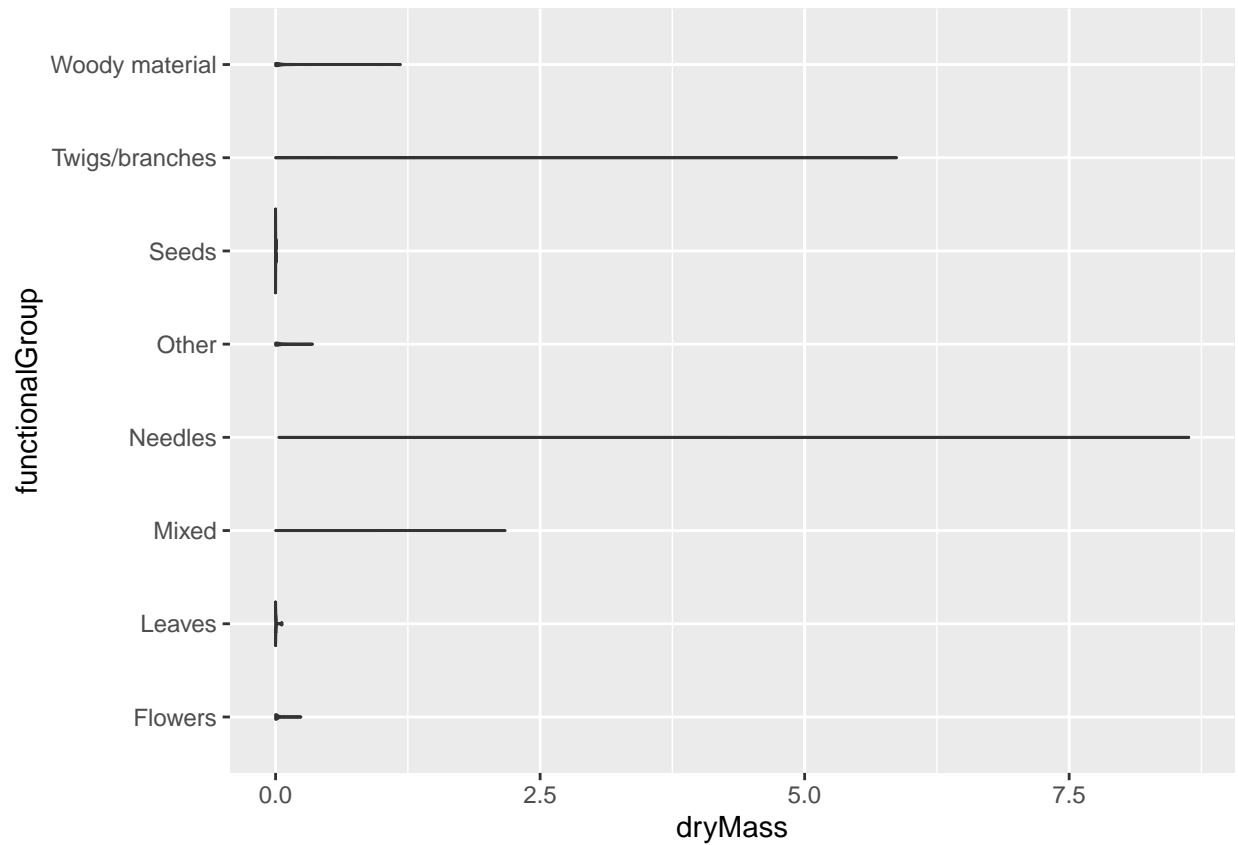


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of `dryMass` by `functionalGroup`.

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

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



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

Answer: The boxplot more clearly displays the IQR and better distinguishes where outliers fall.

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

Answer: Needles and mixed