

Assignment 3: Data Exploration

Natalie Holsclaw

02/02/2025

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Exploration.

Directions

1. Rename this file `<FirstLast>_A03_DataExploration.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change “Student Name” on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Assign a useful **name to each code chunk** and include ample **comments** with your code.
5. Be sure to **answer the questions** in this assignment document.
6. When you have completed the assignment, **Knit** the text and code into a single PDF file.
7. After Knitting, submit the completed exercise (PDF file) to the dropbox in Canvas.

TIP: If your code extends past the page when knit, tidy your code by manually inserting line breaks.

TIP: If your code fails to knit, check that no `install.packages()` or `View()` commands exist in your code.

Set up your R session

1. Load necessary packages (tidyverse, lubridate, here), check your current working directory 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 include the sub-command to read strings in as factors.

```
# Load necessary packages
library(tidyverse)
library(lubridate)
library(here)

# Check working directory
getwd()
```

```
## [1] "/home/guest/EDA_Spring2025"
```

```
# Load datasets
Neonics <- read.csv(
  file = here('Data', 'Raw', 'ECOTOX_Neonicotinoids_Insects_raw.csv'),
  stringsAsFactors = TRUE
)

Litter <- read.csv(
  file = here('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: We might want to know the extent of the impacts of neonicotinoids on insects especially on insects that it may not be targeting as this could influence food webs of ecosystems and other environmental factors.

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: Data on litter and woody debris might give us information about the health of forest ecosystems as these debris are important habitats for certain animals and play a role in nutrient cycling.

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. Litter and woody debris are collected using elevated and ground traps. 2. Traps are collected once a year and debris is separated by size and type. 3. The litter and woody debris are also weighed at collection.

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 “Effect” column, determine the most common effects that are studied. Why might these effects specifically be of interest? [Tip: The `sort()` command is useful for listing the values in order of magnitude...]

```
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 include population, mortality, and behavior. These factors are important when looking at the impacts of neonicotinoids - how they might impact insect populations, mortality, and behavior.

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. [TIP: Explore the help on the `summary()` function, in particular the `maxsum` argument...]

```
summary(Neonics$Species.Common.Name, maxsum = 6)
```

```
##      Honey Bee      Parasitic Wasp Buff Tailed Bumblebee
##           667           285           183
## Carniolan Honey Bee      Bumble Bee      (Other)
##           152           140           3196
```

Answer: This group of insects includes all pollinators which are vital for agriculture and plant reproduction.

8. Concentrations are always a numeric value. What is the class of `Conc.1..Author.` column in the dataset, and why is it not numeric? [Tip: Viewing the dataframe may be helpful...]

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

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

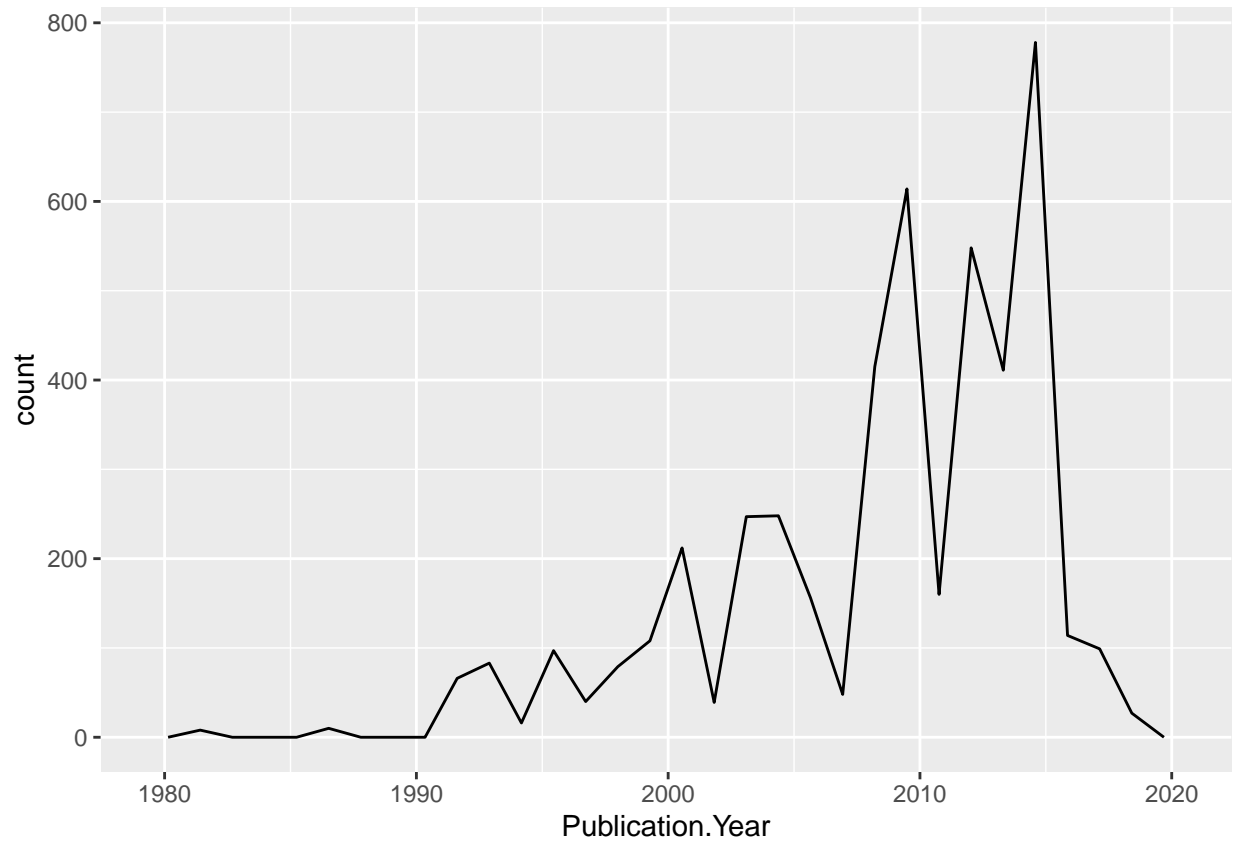
Answer: This column is classified as ‘factor’ probably because it contains inequality symbols.

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))
```

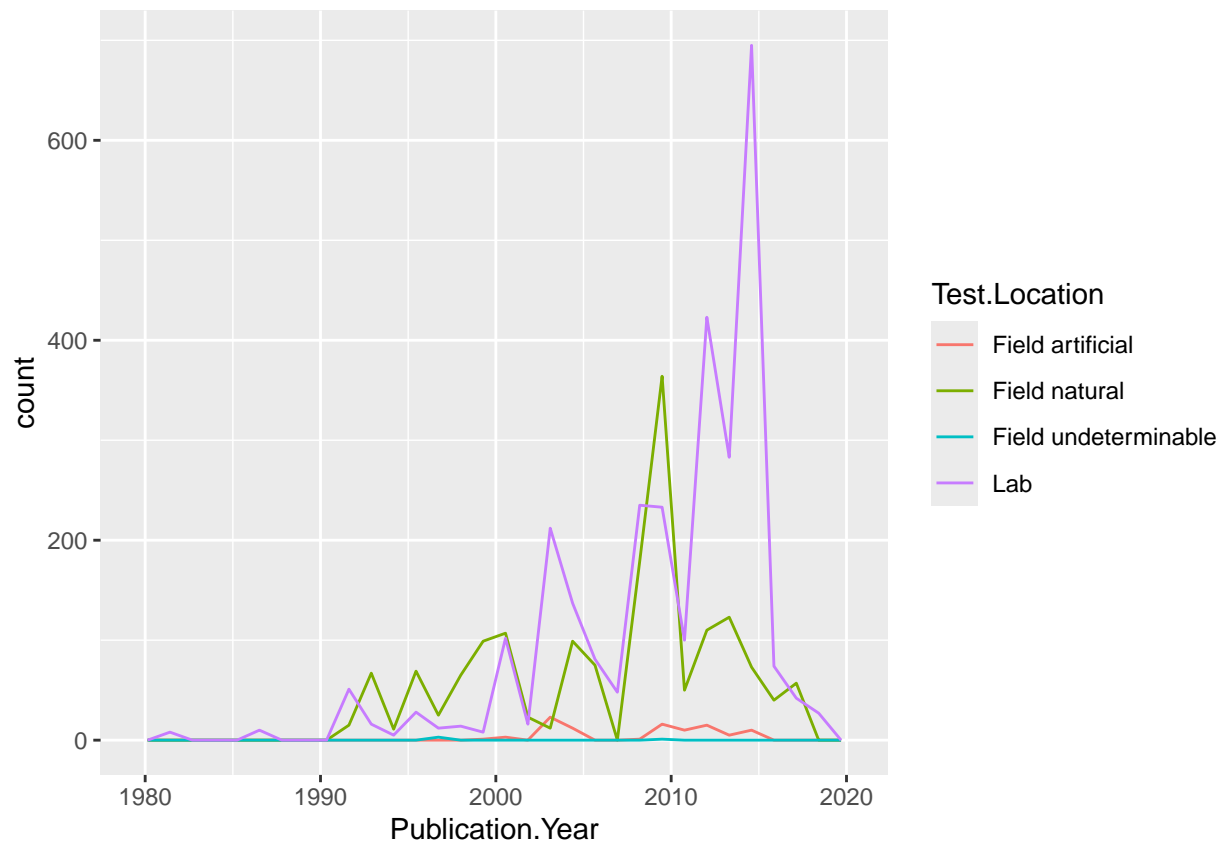
'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.



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'.



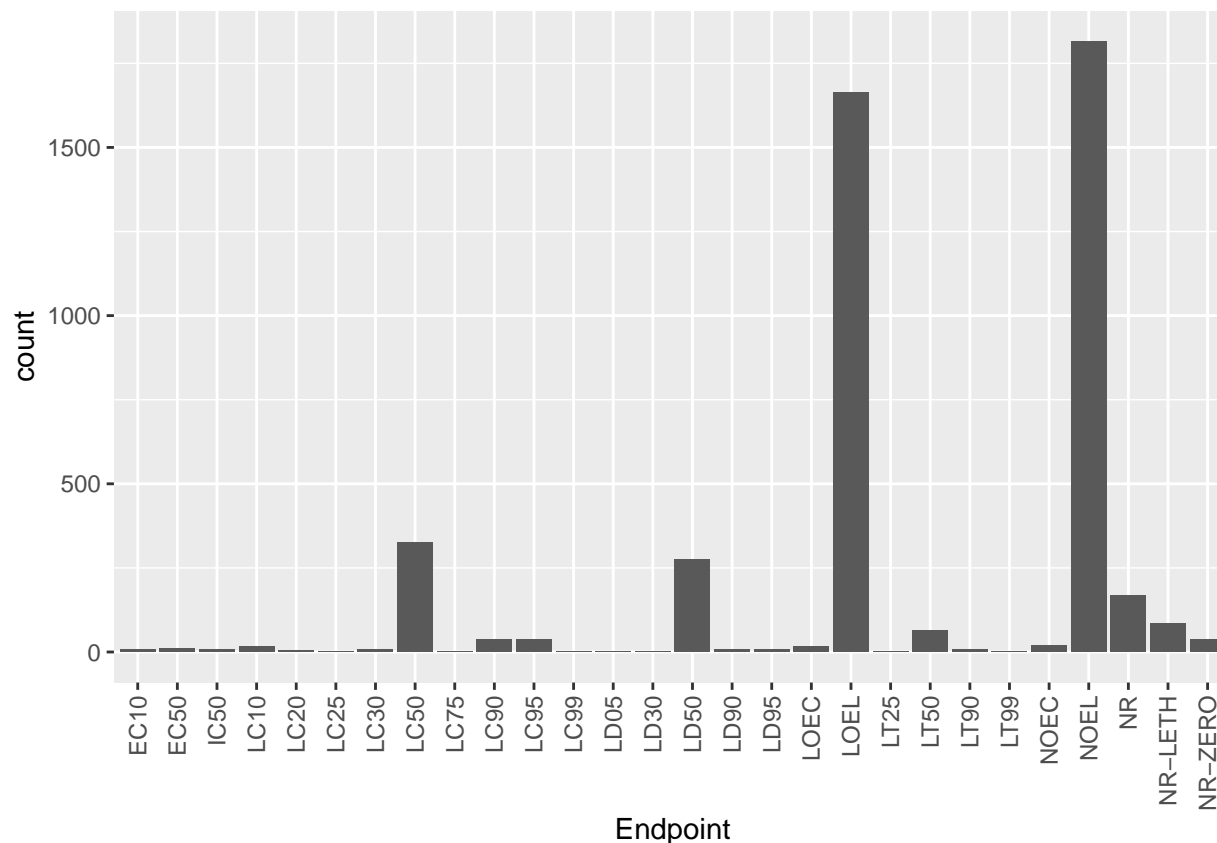
Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: The most common test location is a lab followed by “field natural.” Between 2010 and 2015 lab studies increased significantly while others decreased.

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.

[**TIP:** Add `theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))` to the end of your plot command to rotate and align the X-axis labels...]

```
ggplot(data = Neonics, aes(x = Endpoint)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: The two most common endpoints are NOEL and LOEL or no-observable-effect-level and lowest-observable-effect-level respectively.

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.

```
# Original class of 'collectDate' column
class(Litter$collectDate)
```

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

```
# Conver 'collectDate' to date format - check class
Litter$collectDate <- ymd(Litter$collectDate)
class(Litter$collectDate)
```

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

```
# Specific dats litter was sampled in August 2018
unique(Litter$collectDate)
```

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

13. Using the `unique` function, determine how many different 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
```

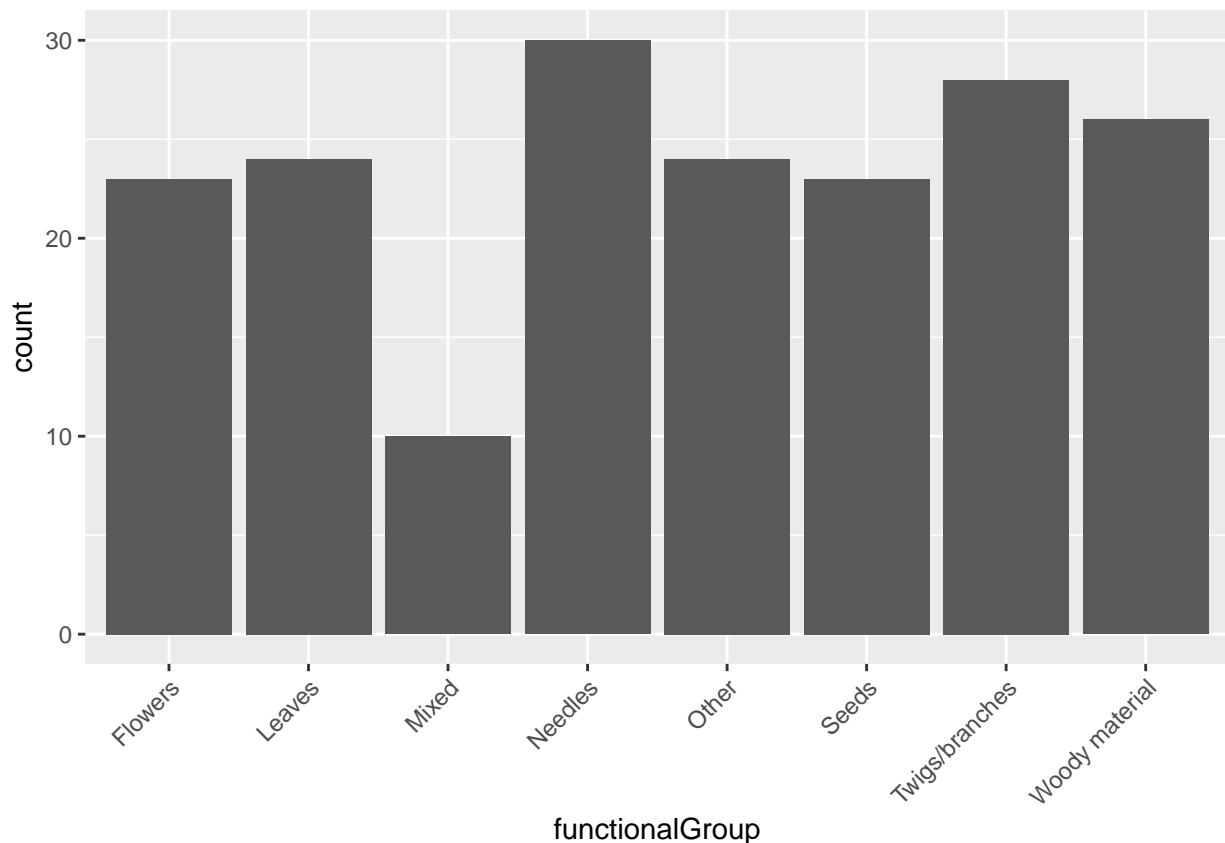
```
summary(Litter$plotID)
```

```
## 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 functions tells us there are 12 ‘levels’ or 12 unique plot IDs and the summary function tells us how many samples fall into each plot ID category.

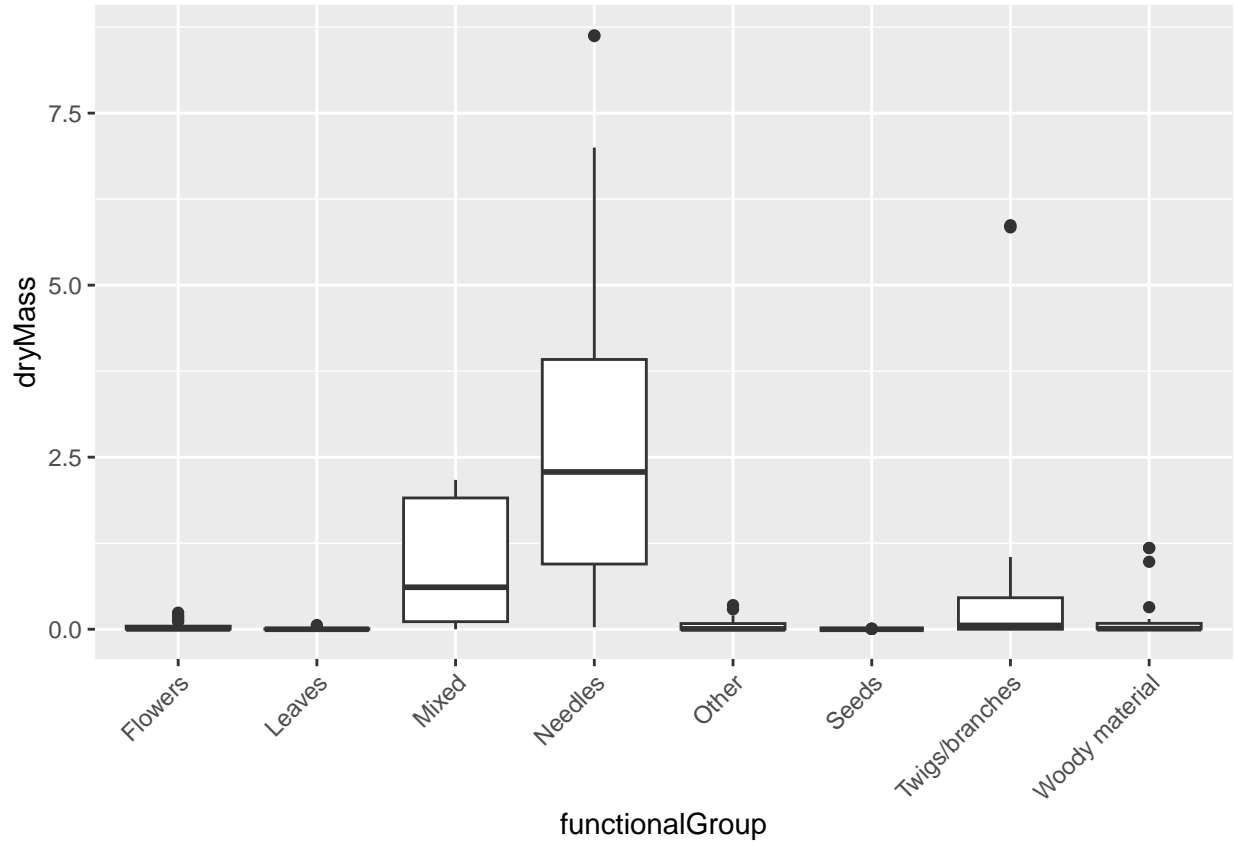
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(data = Litter, aes(x = functionalGroup)) +  
  geom_bar() +  
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1))  
)
```

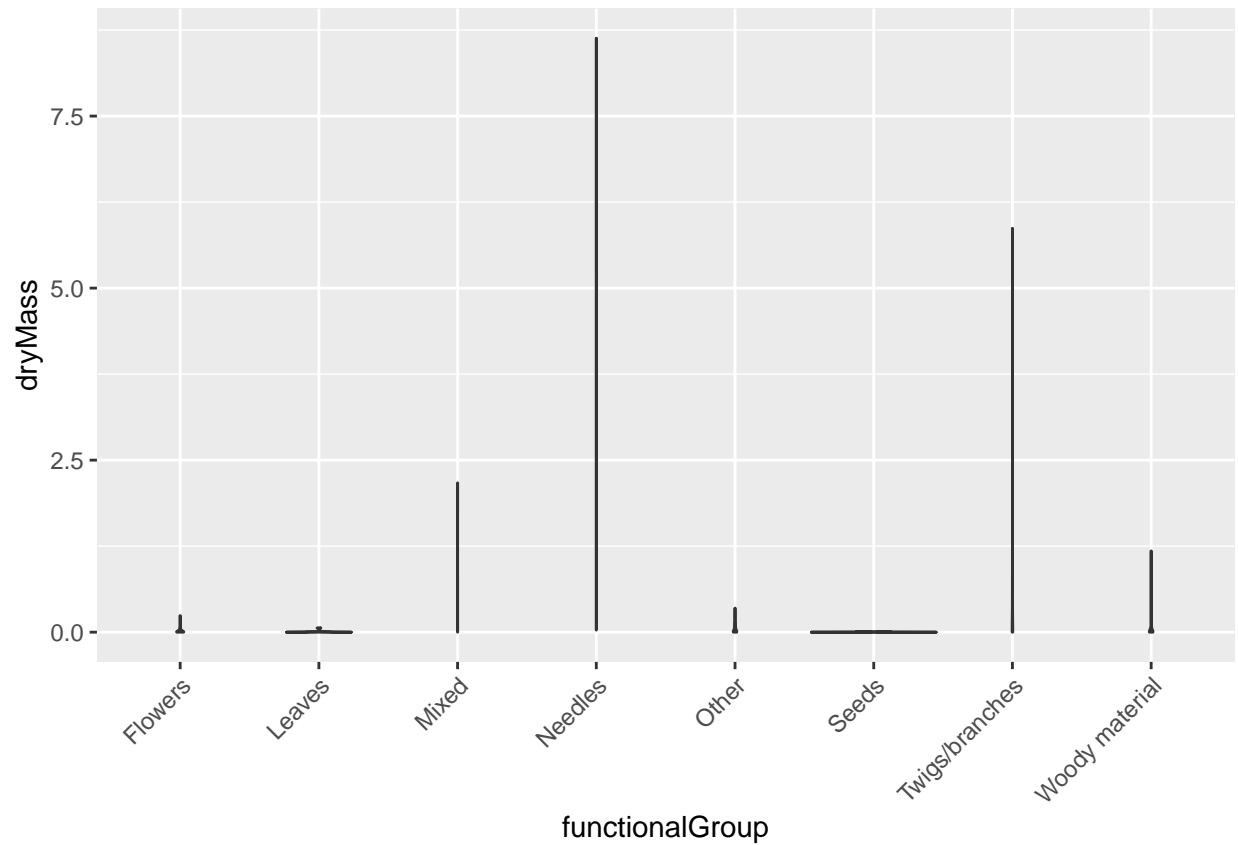


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of `dryMass` by functional-Group.

```
ggplot(Litter)+  
  geom_boxplot(aes(x = functionalGroup, y = dryMass))+  
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1)  
        )
```



```
ggplot(Litter)+  
  geom_violin(aes(x = functionalGroup, y = dryMass))+  
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```

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

Answer: Since our data on the x-axis is categorical, the violin plot just shows up as lines so the boxplot is more effective.

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

Answer: Needles and mixed litter tend to have the highest biomass.