Assignment 3: Data Exploration

Rachel Williams

Fall 2024

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 subcommand to read strings in as factors.

```
#Loading necessary packages
library(tidyverse); library(lubridate); library(here)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
## here() starts at /home/guest/EDE_Fall2024
```

```r
#Reading in Neonics and Litter data
Neonics <- read.csv(
  file=here('Data','Raw','ECOTOX_Neonicotinoids_Insects_raw.csv'),
  stringsAsFactors = T
)

Litter <- read.csv(
  file=here('Data','Raw','NEON_NIWO_Litter_massdata_2018-08_raw.csv'),
  stringsAsFactors = T)
```

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 be interested on the effect of neonicotinoids because while they are targetted at pests,they also harm helpful species such as bees. Studying it's effects on bugs could tell us which ones are unintentionally harmed.

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 in forests might be of interest because some species rely on it for nutrients or shelter. It has an important role in nutrient cycling and the carbon cycle, and can also affect streamflow.

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 traps were used to collect data. The debris was collected from the traps by plant functional type. 2. Litter waws collected from elevated traps and fine debris was collected from ground traps. 3. Mass was collected for samples based on their functional group. These groups included leaves, needles, seeds, flowers, twigs, and woody material.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
dim(Neonics) #Shows number of rows and columns.
```

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

```r
#There are 4623 rows and 30 columns.
```

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...]

```r
sort(summary(Neonics$Effect), decreasing = T)
```

```
##       Population        Mortality         Behavior Feeding behavior
##             1803             1493              360              255
##     Reproduction      Development        Avoidance         Genetics
##              197              136              102               82
##        Enzyme(s)           Growth       Morphology    Immunological
##               62               38               22               16
##     Accumulation     Intoxication     Biochemistry          Cell(s)
##               12               12               11                9
##       Physiology        Histology       Hormone(s)
##                7                5                1
```

```r
#decreasing was used to sort from highest to lowest
```

Answer:The most common effects are population and mortality. These might be of interest because they show if a population is increasing or decreasing over time. This shows which species are the most threatened and which might need protection.

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...]

```r
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
```

```r
#Maxsum lists only the top 6 values.
```

Answer: The most common species were honey bees, parasitic wasps, buff tailed bumblebees, Carolinan Honey Bee, Bumble Bee, and other. All of these species are bees or wasps. This is important because they are pollinators that are needed for plant health. THe fact that they are dying more than pests is cause for concern.

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"
```

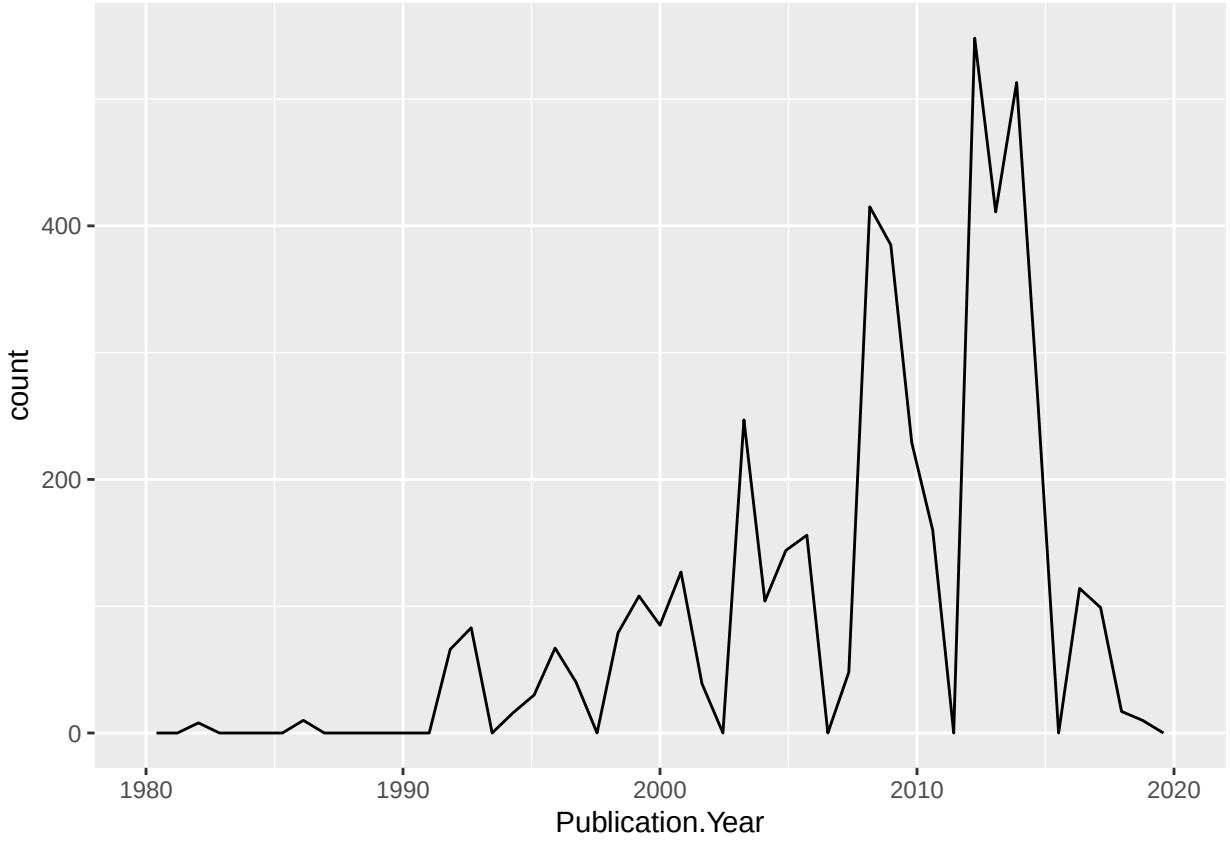
Answer: THe class of this column is factor. It isn't numeric because it is paired with units in Conc.1.Units.Author.

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=50) +
  scale_x_continuous(limits = c(1980 , 2020))
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_path()`).
```

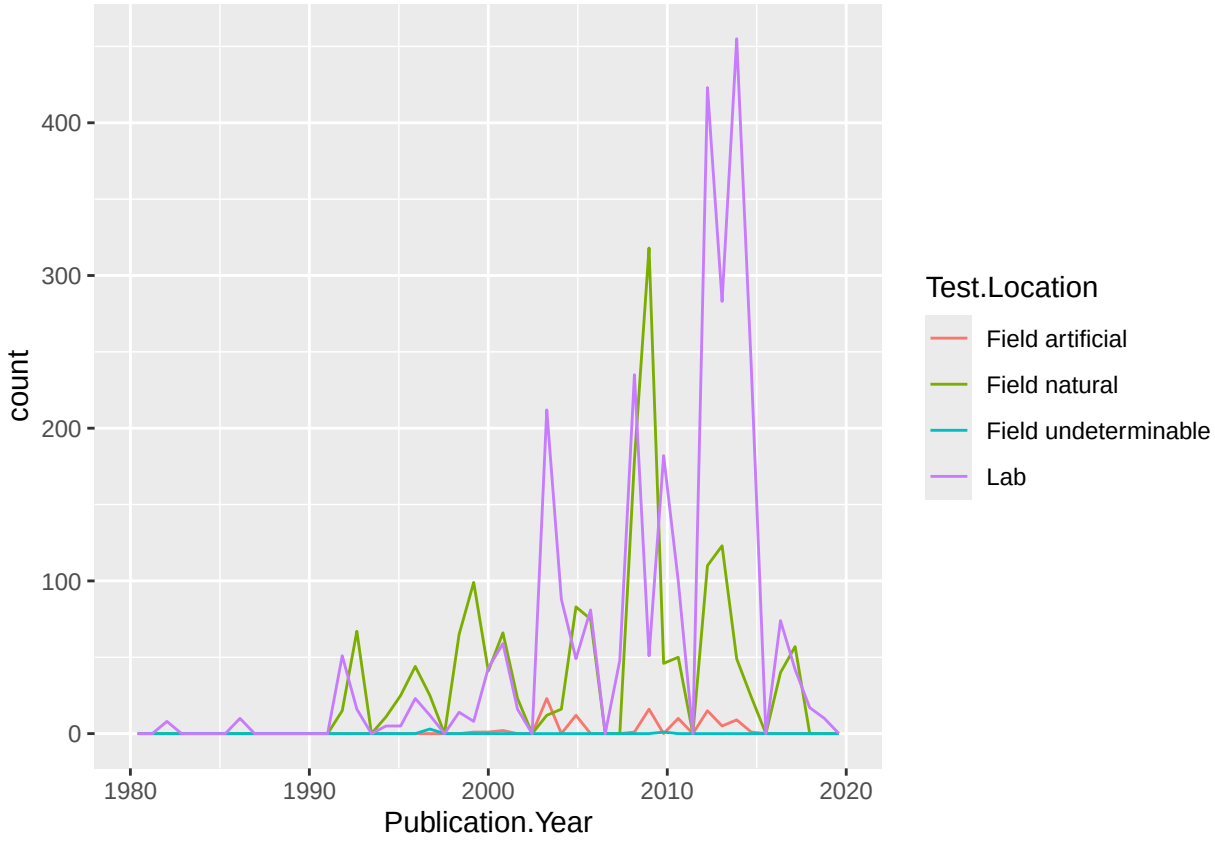



```
#Scale must be continuous not discrete; I chose the limits based on the highest
#and lowest dates in the data.
```

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=50)+ scale_x_continuous(limits = c(1980 , 2020))
```

```
## Warning: Removed 8 rows containing missing values or values outside the scale range
## (`geom_path()`).
```




```
#Whatever part of the data color is will be differentiated.
```

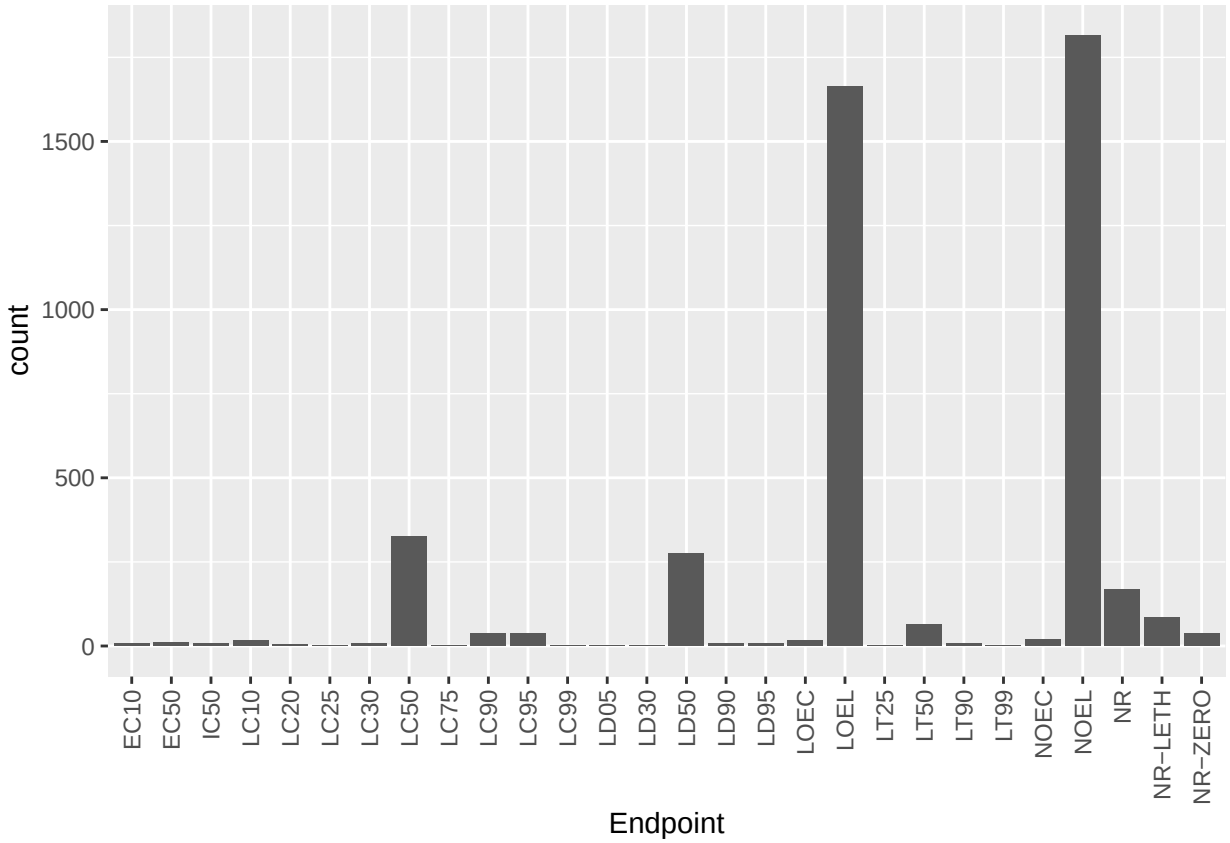
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locactions are Lab and Field natural. Lab becomes more popular over time and had the highest usage. The natural field was used often until usage decreased dramatically around 2010.

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 end points are LOEL and NOEL. NOEL means no observable effect level and LOEL means lowest observable effect level.

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)#factor, not data
```

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

```
Litter$collectDate <- ymd(Litter$collectDate)# Use function values in correct order
class(Litter$collectDate)
```

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

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

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

```
#The dates litter was collected were 08/02/2018 and 08/30/2024
```

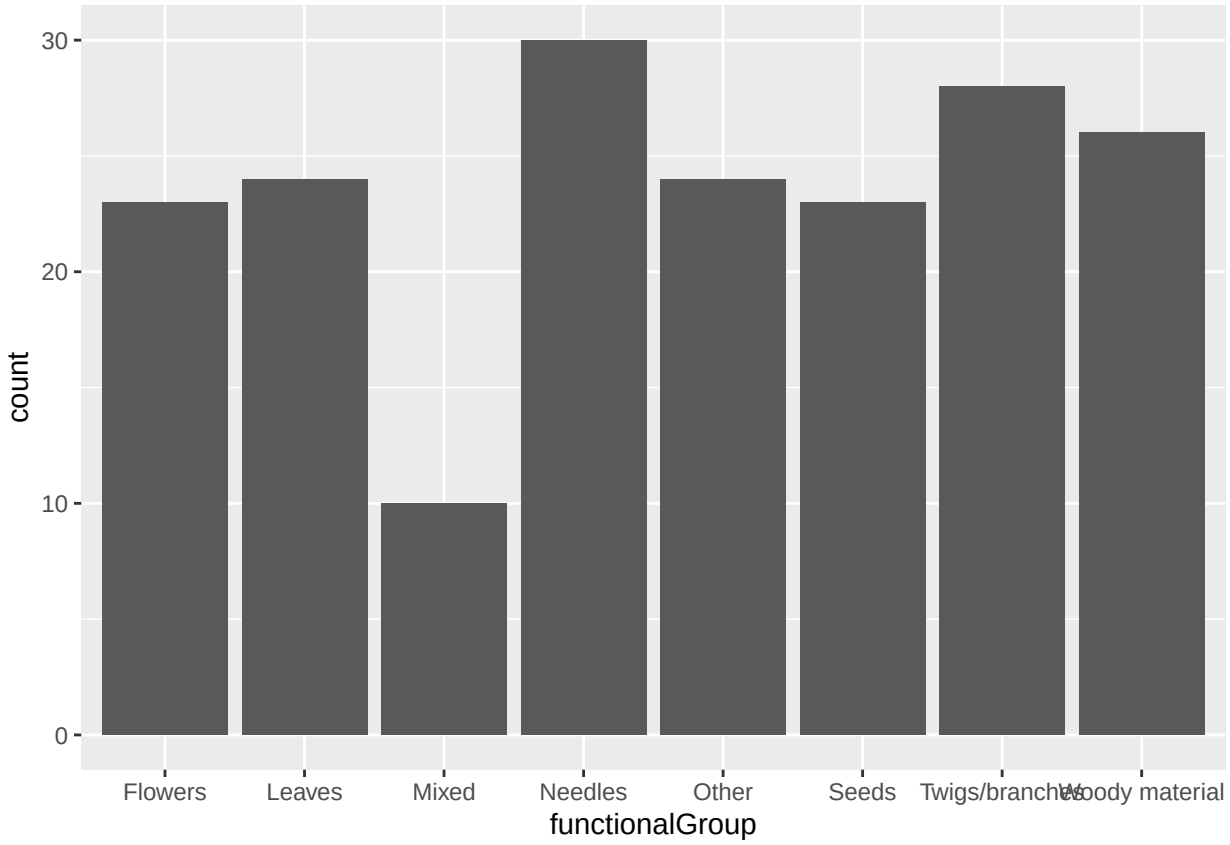
13. Using the `unique` function, determine how many different plots were sampled at Niwot Ridge. How is the information obtained from `unique` different from that obtained from `summary`?

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

Answer: 12 different plots were sampled at Niwot Ridge. The information from unique is different because it only includes the number of unique values. Summary also includes how many instances there were of each value.

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.

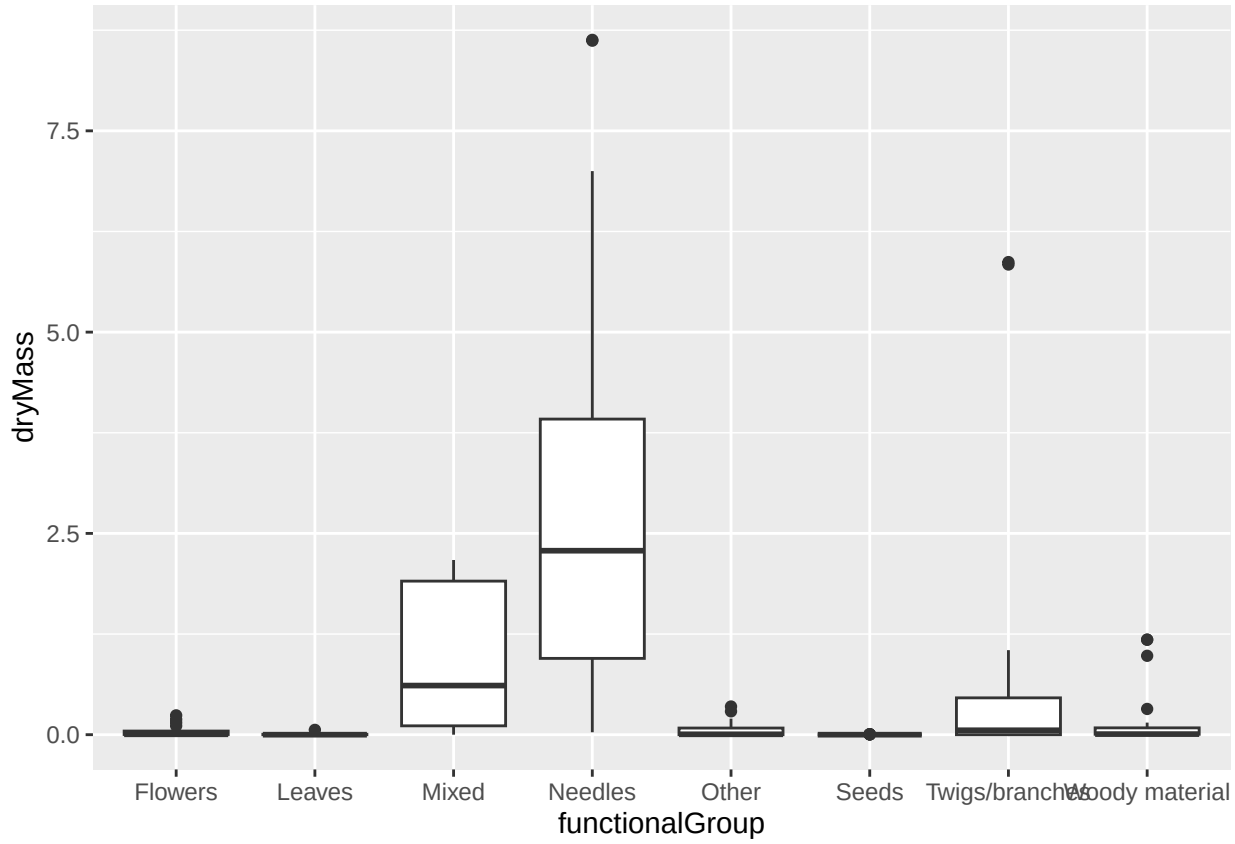
```
ggplot(data = Litter, aes(x = functionalGroup)) +
  geom_bar()
```



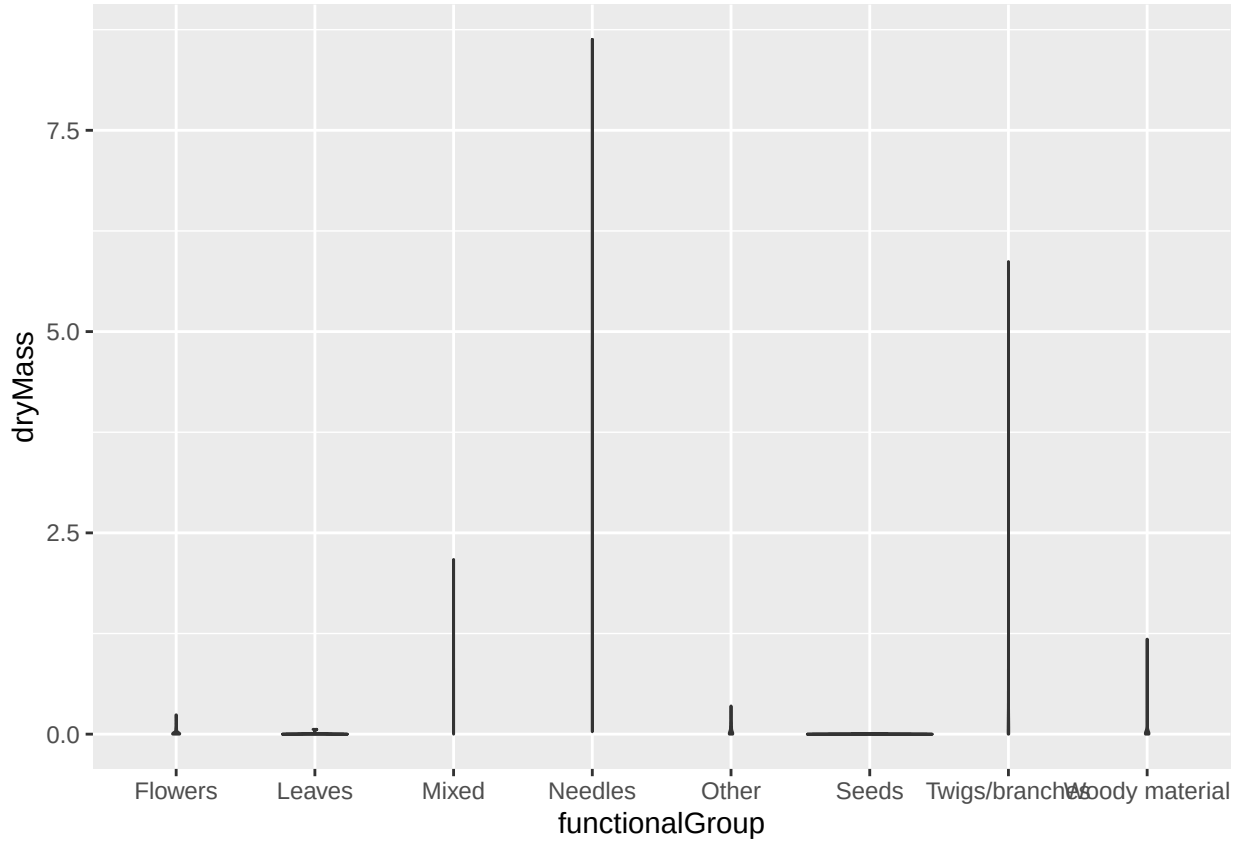

```
#This bar plot showed less mixed than other groups.
```

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,
                 group = cut_width(functionalGroup, 1)))
```




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



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

> Answer: In this case the box plot is more effective because it shows a clearly defined median, quartiles, and maximum point for the groups with the most mass. THe violin plot is too thin to be easily read.

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

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