Assignment 3: Data Exploration

Kendall Barton

Fall 2023

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 Sakai.

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. Check your working directory, load necessary packages (tidyverse, lubridate), 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.

```r
getwd() #wd is set to the EDE_Fall2023 folder!
```

```
## [1] "/Users/kendallbarton/Downloads/EDE_Fall2023"
```

```r
library(tidyverse)
library(lubridate)
Neonics <- read.csv('./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv',stringsAsFactors = TRUE)
#datasets are already downloaded in EDE_Fall2023 folder
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: Neonicotinoids are insecticides, which are supposed to be toxic to insects. Ecotoxicology of neonicotinoids on insects is interesting in order to confirm that neonicotinoids are working the way they are intended, or to see the consequences of neonicotinoids on non-target insect species.

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: It could be used as an indirect measure of primary productivity for a site.

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 is considered to be material dropped from the forest canopy that is <2cm in end diameter. 2. Litter and debris is caught in a 0.5 cmˆ2 PVC square with a mesh basket ~80cm above the ground. 3. Sampling occurs at the same locations until it is no longer possible, then that location and ID are retired.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) #dimensions of Neonics (4623 x 30)
```

```
## [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?

```
sort(summary(Neonics$Effect), decreasing = TRUE) #sort effects most to least common
```

```
##       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
```

Answer: Studies examining neonicotinoids and insect mortality and population are the most common, which makes sense because neonicotinoids kill insects.

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: The `sort()` command can sort the output of the summary command...]

```r
sort(summary(Neonics$Species.Common.Name), decreasing = TRUE)[1:6] #top 6 most common species
```

```
##                (Other)              Honey Bee         Parasitic Wasp
##                    670                    667                    285
## Buff Tailed Bumblebee   Carniolan Honey Bee             Bumble Bee
##                    183                    152                    140
```

Answer: The most commonly studied species (besides "Other") are all bees except for one species of wasp. This is probably because these insects are pollinators. Pollinators are beneficial so people don't want neonicotinoids to kill them, unlike the pests that the pesticides are targeting.

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?

```r
class(Neonics$Conc.1..Author.) #find class
```

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

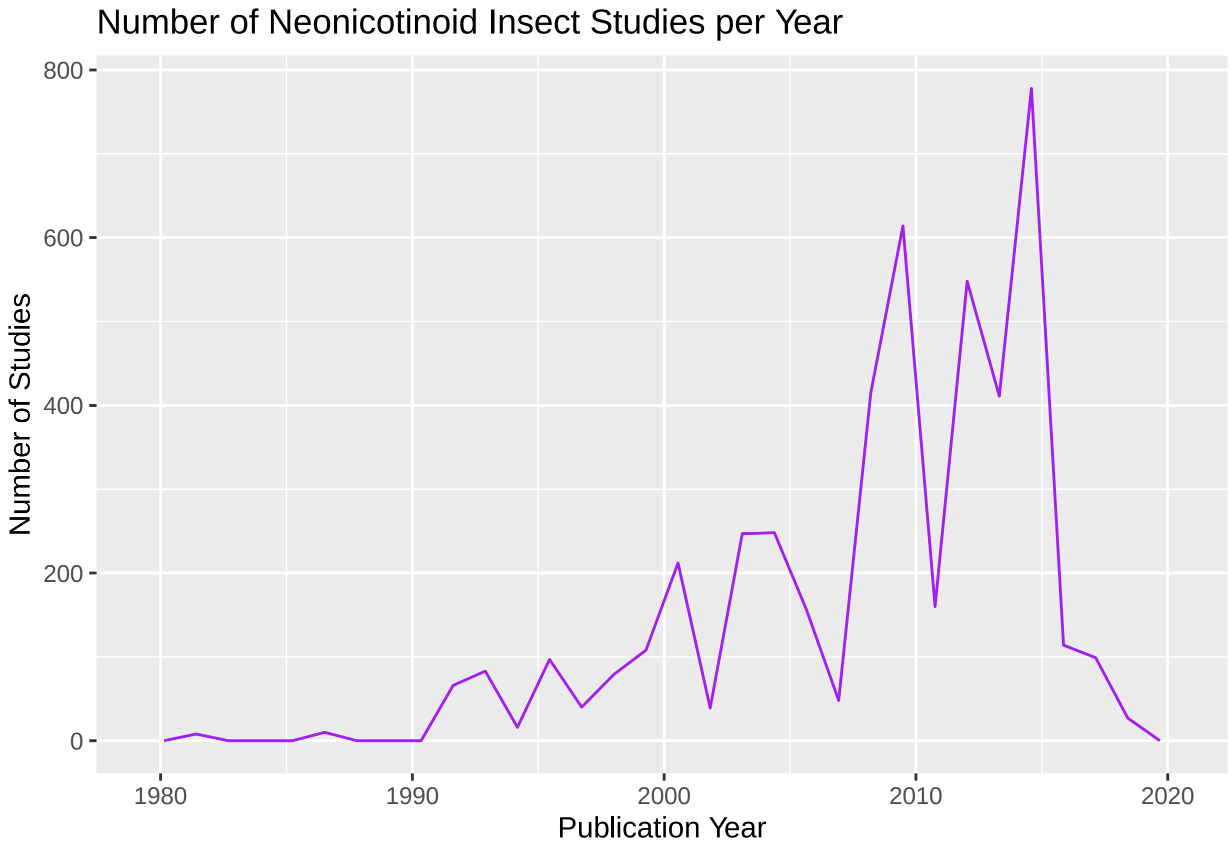
Answer: It's a factor because when we imported the data, we said that stringsAsFactors = TRUE.

Explore your data graphically (Neonics)

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

```r
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), color = "purple") + #I just like purple
  xlab("Publication Year") + ylab("Number of Studies") + #adding axis labels
  ggtitle("Number of Neonicotinoid Insect Studies per Year") #why not have a title?
```

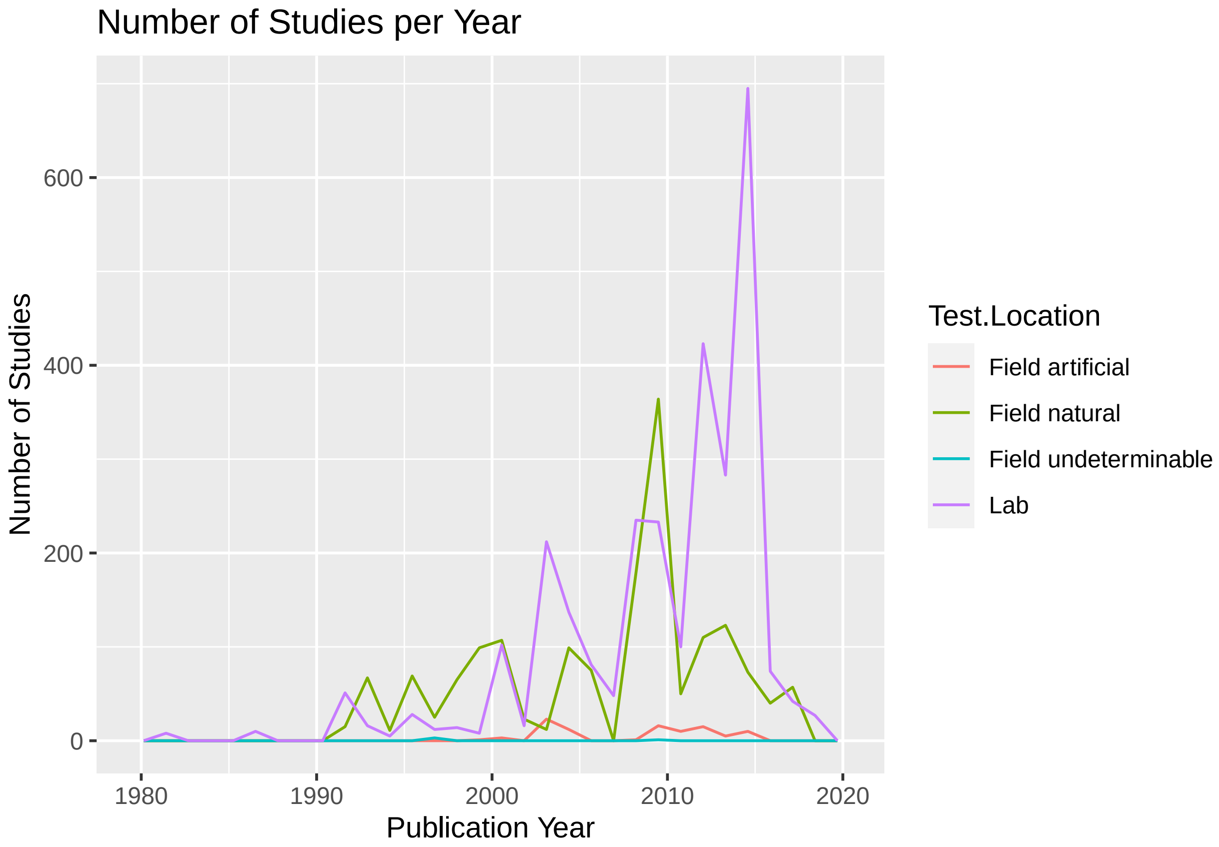
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```r
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location)) +
  #this graph has different lines for different test locations
  xlab("Publication Year") + ylab("Number of Studies") +
  ggtitle("Number of Studies per Year")
```

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



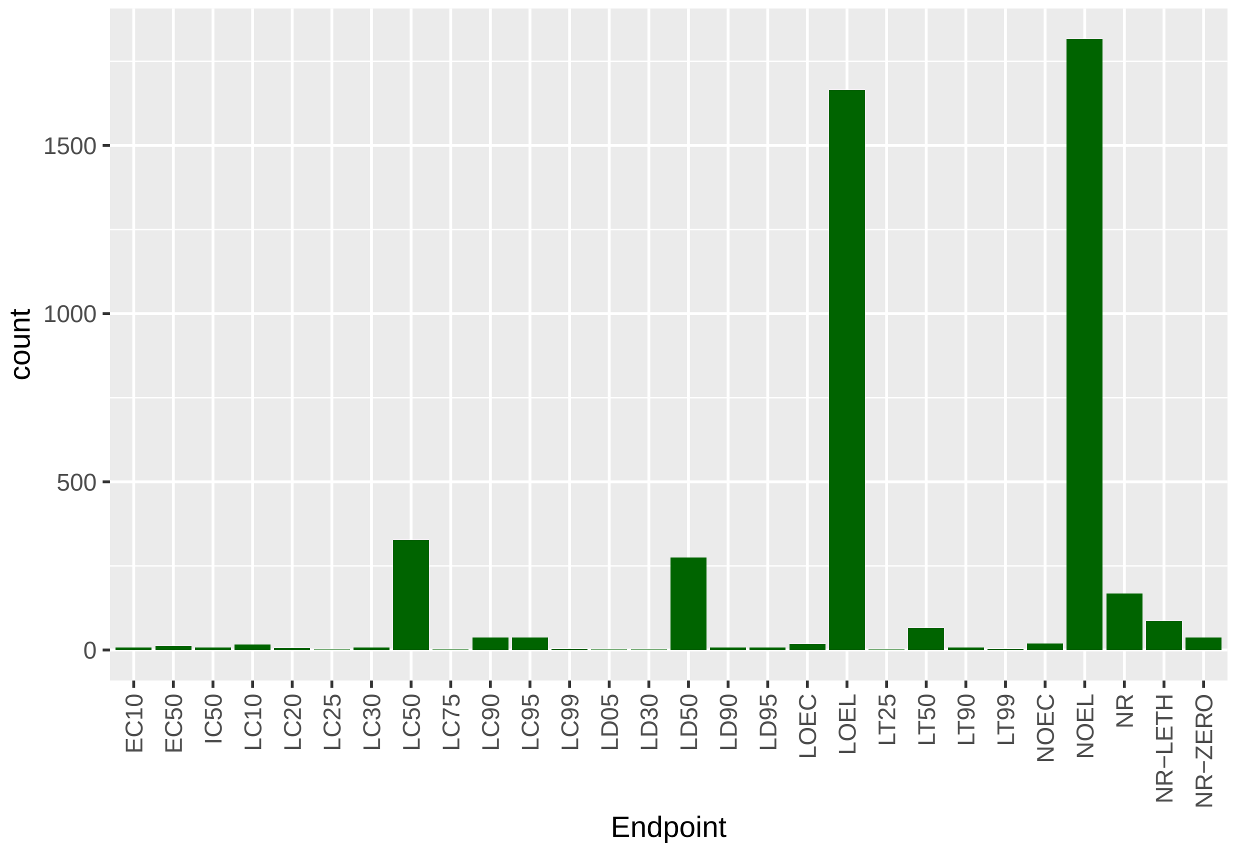
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: Lab and field natural are the two most common test locations for studies after 1990. Until 2000, field natural studies were more numerous. After 2000, lab has been more popular for the most part, with a couple of especially big spikes 2010-2015.

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(Neonics) +
  geom_bar(aes(Endpoint), fill = "darkgreen") + #I like green too
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) #I like the rotation!
```



Answer: The most common end points are LOEL (lowest-observable-effect-level) and NOEL (no-observable-effect-level). LOEL is the lowest dose that produces significantly different effects from the control. NOEL is the highest dose that doesn't produce effects significantly different from the control.

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) #find original class
```

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

```
Litter$collectDate <- ymd(Litter$collectDate) #change to date class
print(class(Litter$collectDate)) #confirm date class
```

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

```
print(unique(Litter$collectDate)) #find sample dates
```

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

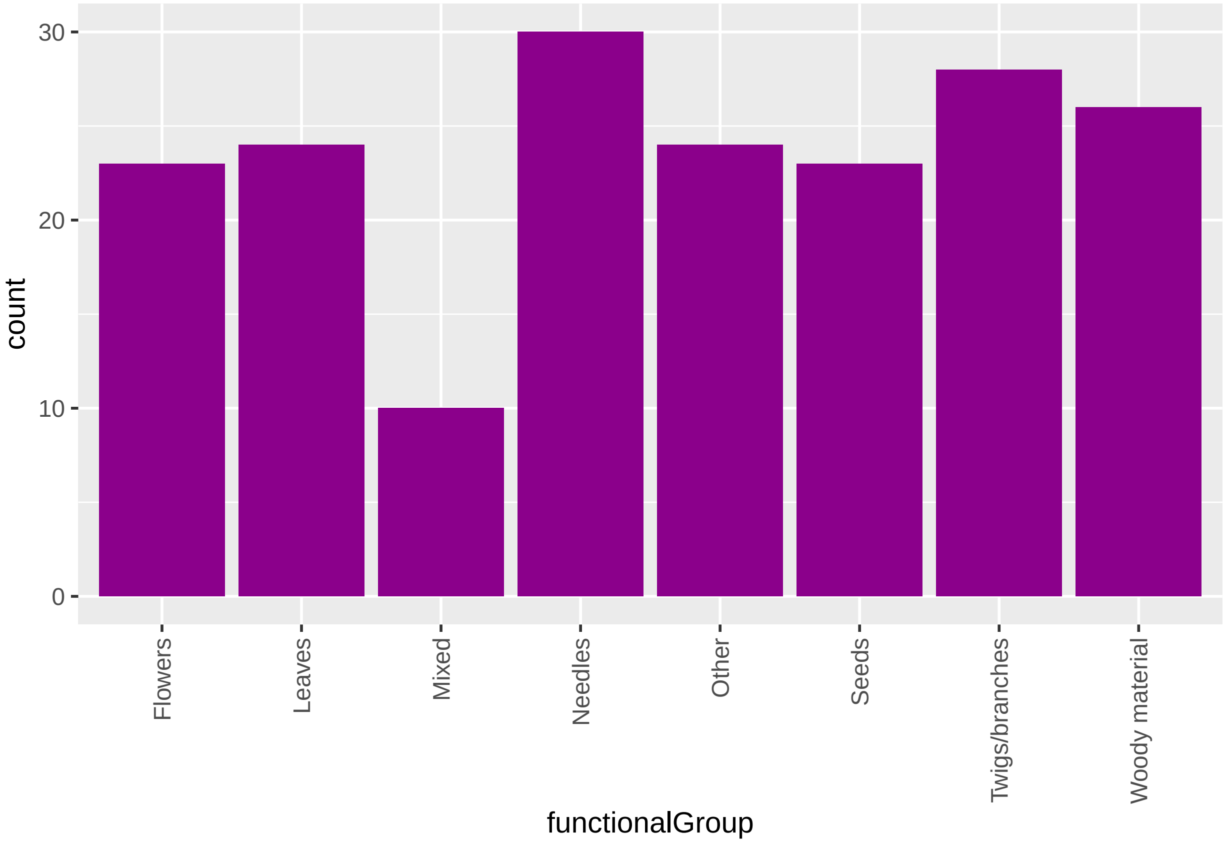
```
print(length(unique(Litter$namedLocation))) #number of unique sites
```

```
## [1] 12
```

Answer: 'unique' lists the unique variables, while 'summary' lists the unique variables and how many times they each appear.

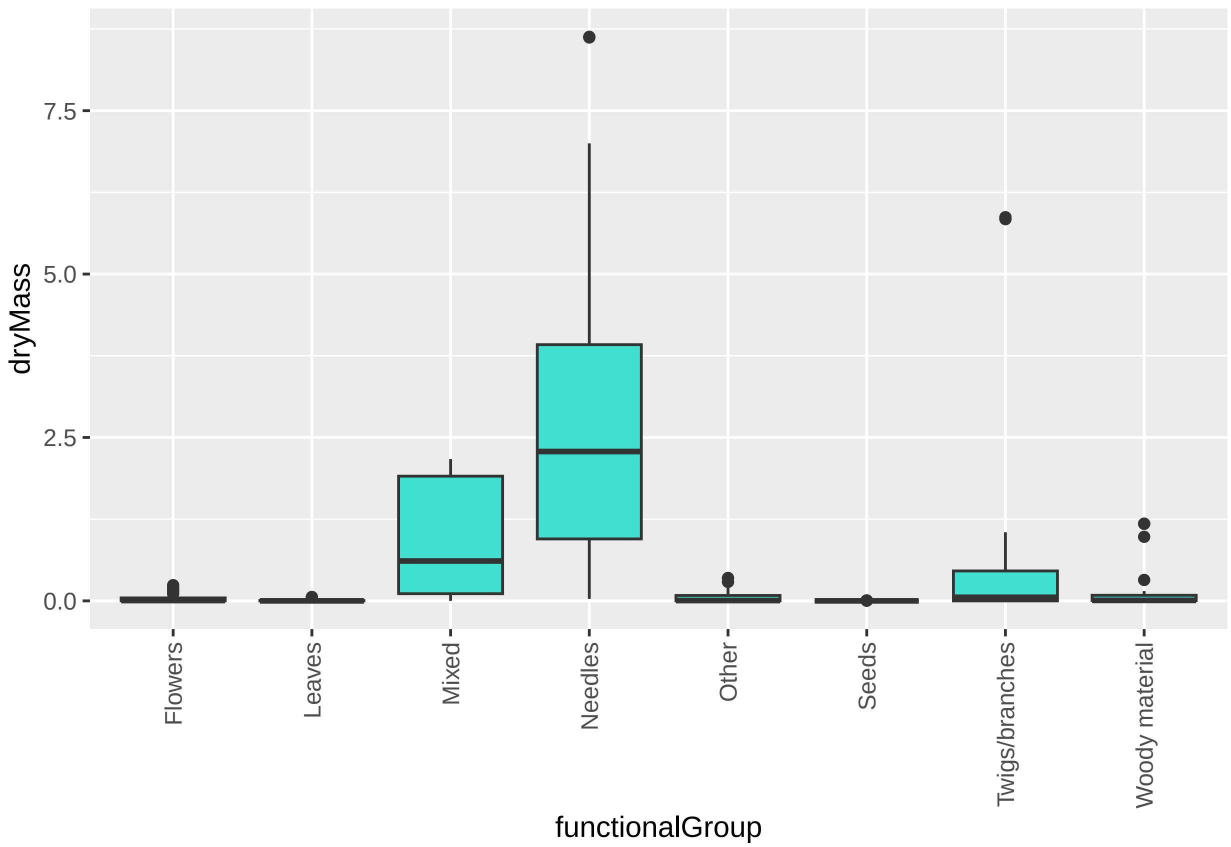
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter) +
  geom_bar(aes(functionalGroup), fill = "darkmagenta") + #fav color so far
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) #easier to see x axis labels
```

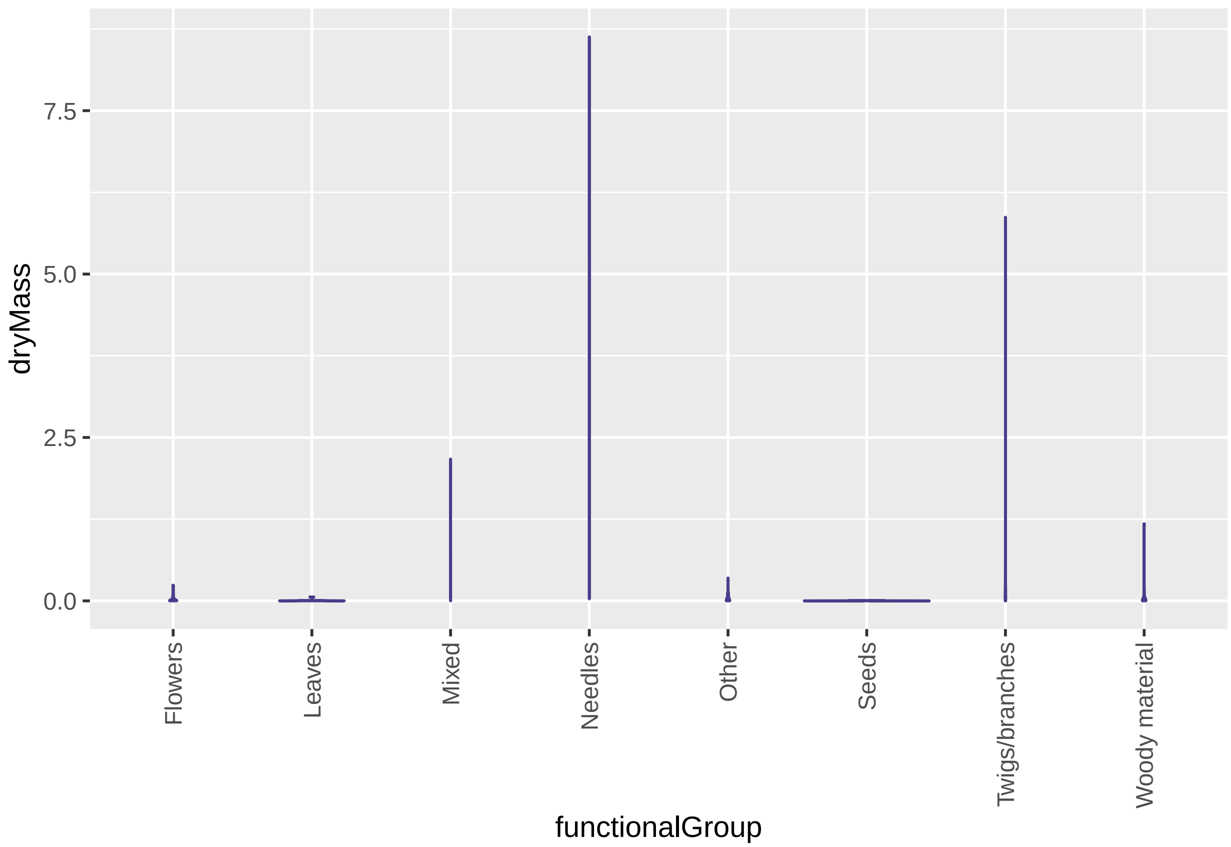


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(Litter) +
  geom_boxplot(aes(x=functionalGroup, y=dryMass), fill = "turquoise") +
  #dryMass as a function of functionalGroup
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



```r
ggplot(Litter) +
  geom_violin(aes(x=functionalGroup, y=dryMass), color = "darkslateblue") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



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

> Answer: The boxplot is better because the violin plot looks like straight lines for the most part and isn't clearly understandable or informative. I assume that the violin plot looks this way because there are too few data points.

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

> Answer: Needles are the clear frontunner for the highest biomass.