BrookHemphill_A03_DataEploration.Rmd

Brook Hemphill

Spring 2024 01/28/2024

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 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.

```
# check working directory
getwd()
```

```
## [1] "/home/guest/EDA_Spring2024_New"
```

```
# Load packages
library(tidyverse)
library(dplyr)
library(ggplot2)
```

```r
#Load datasets
#ECOTOX_neonicotinoid <-
#read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",stringsAsFactors = TRUE)

#Niwot_Ridge_NEON <-
#read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv",stringsAsFactors = TRUE)

Niwot_Ridge_NEON <-
  read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv",stringsAsFactors = TRUE)

ECOTOX_neonicotinoid <-
  read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_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 are interested in understanding the ecotoxicilogy of neonicotinoids on insects due to their high toxicity in insects, which can be lethal. Suscepticle insects, including pollinators can suffer from directly consuming this substance, which can impact the food web leading to cascading effects in ecosystems.

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: We are interested in litter and woody debris on the forest floor for several reasons. These components of the forest floor are indicators for forest health and biodiversity (it provides food for creatures and shelter), they are a crucial role in nutrient cycling and soil formation, and important carbon pools.

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. ground traps are sample one time per year 2. sites with deciduous vegetation/during winter season, sampling of elevated traps for litter/fine woody debris may be discontinued 3. measured areas with woody vegetation >2m tall

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
# show rows and columns of dataset using Dim()

dim(ECOTOX_neonicotinoid) # 4623 rows 30 columns
```

```
## [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?

```r
#use summary() to look at effects
neonic_effect <- summary(ECOTOX_neonicotinoid$Effect)
```

Answer: The most common effects studied are population (1803). These effects may specifically be of interest to understand how the neonicotnoid is impacting the population of insects and abudance of insect species. For example, we could look at insect population data from year 2020 and year 2021 to determine if the populations were decreasing over time.

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
#use the summary function to get the common species in the dataset
species_summary <- summary(ECOTOX_neonicotinoid$Species.Common.Name)

#use the sort function to sort the most commonly studied species

top_species <- sort(species_summary, decreasing = TRUE)[1:6]
print(top_species)
```

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

Answer:The 6 most commonly studied insects were pollinators, with majority being bees. These insects might be of interest because they provide valuable ecosystem services (pollinating species, food sources for birds and other species) and are incredibly important to our agricutlural industry by pollinating crops.

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
#determine class of concentrations in neonic dataset
concentrations_neonic <- class(ECOTOX_neonicotinoid$Conc.1..Author.)
```

Answer:The class is factor. The class was not numeric because R may have interpreted the data in the column categorical or nominal rather than numeric.

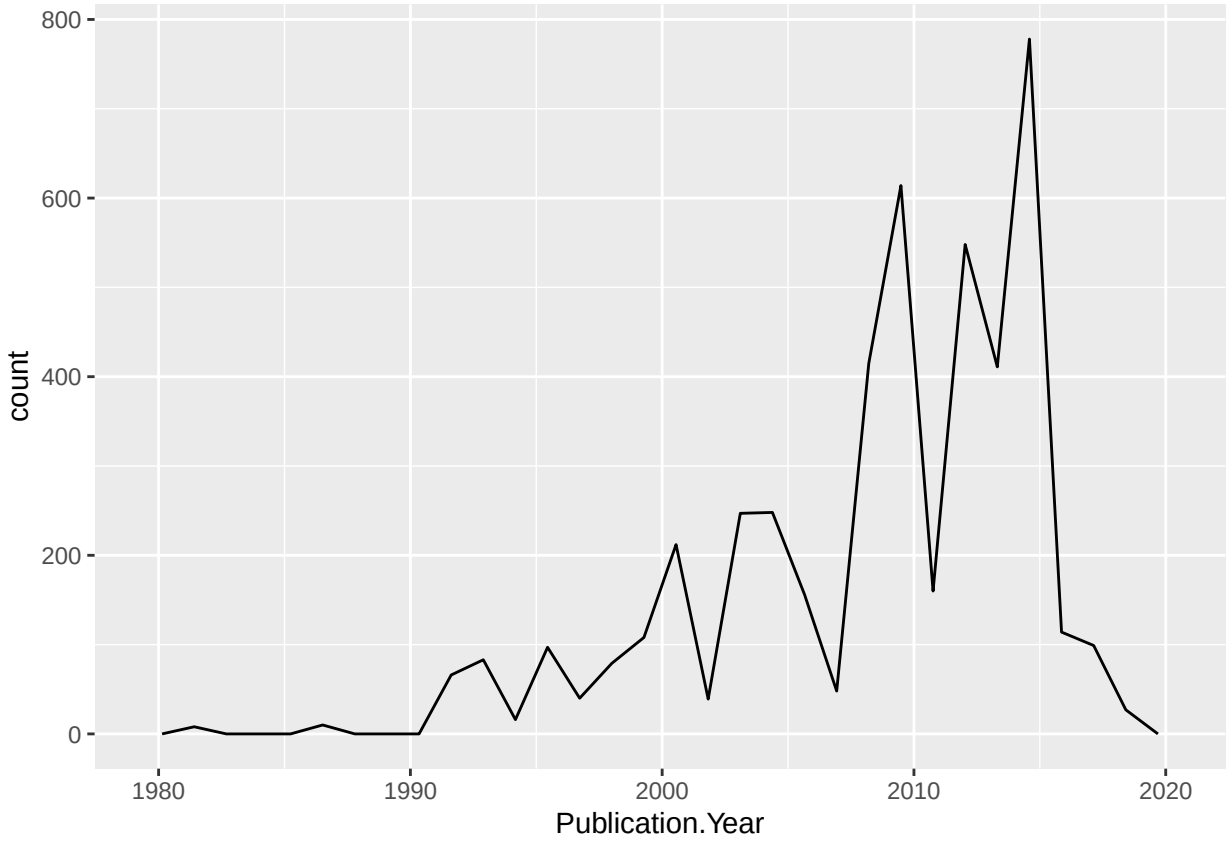
Explore your data graphically (Neonics)

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

```
# plot

ggplot(ECOTOX_neonicotinoid, aes(x = Publication.Year)) +
  geom_freqpoly()
```

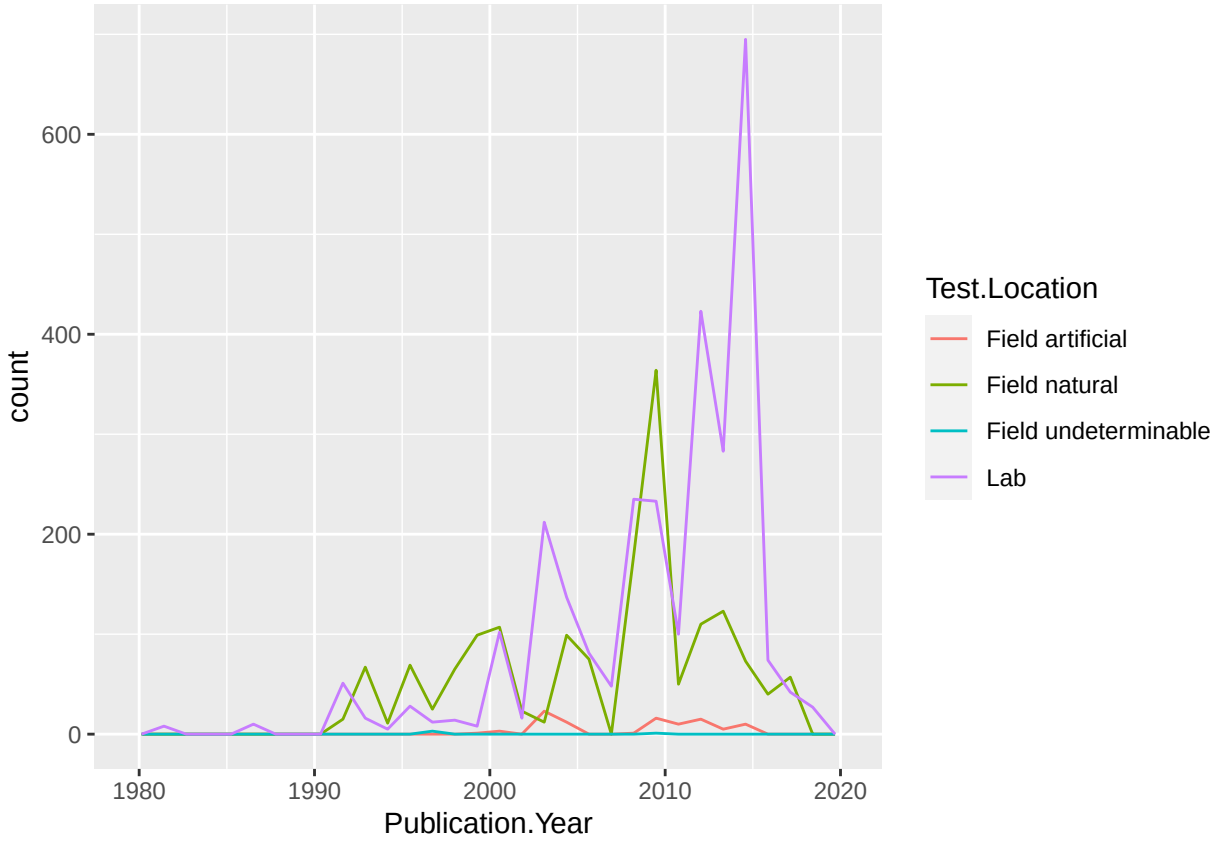
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
ggplot(ECOTOX_neonicotinoid,
       aes(x = Publication.Year, color = Test.Location)) +
  geom_freqpoly() #inside adding colors to the line
```

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



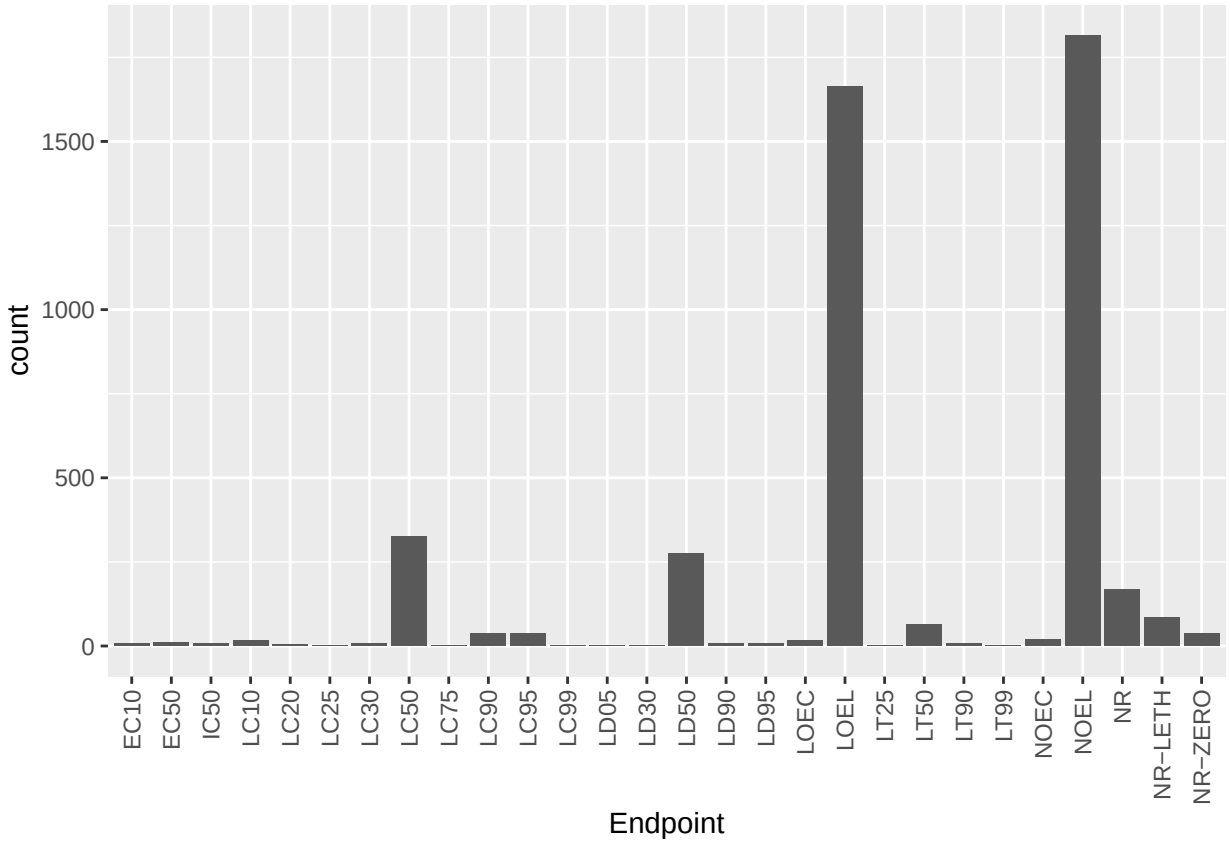
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are in the lab and out in the field. Intially, field test locations were slightly greater and over time the lab test locations were greater.

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



Answer:LOEL and NOEL. LOEL: Terrestrial LOEL Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC). NOEL: Terrestrial NOEL 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)

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.

```
#determine class of collectdate
class(Niwot_Ridge_NEON$collectDate)
```

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

```
Niwot_Ridge_NEON$collectDate <-
  as.Date(Niwot_Ridge_NEON$collectDate, format = "%Y-%m-%d")
unique(Niwot_Ridge_NEON$collectDate[format(Niwot_Ridge_NEON$collectDate, "%Y-%m") == "2018-08"])
```

```
## [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`?

```
unique(Niwot_Ridge_NEON$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
```

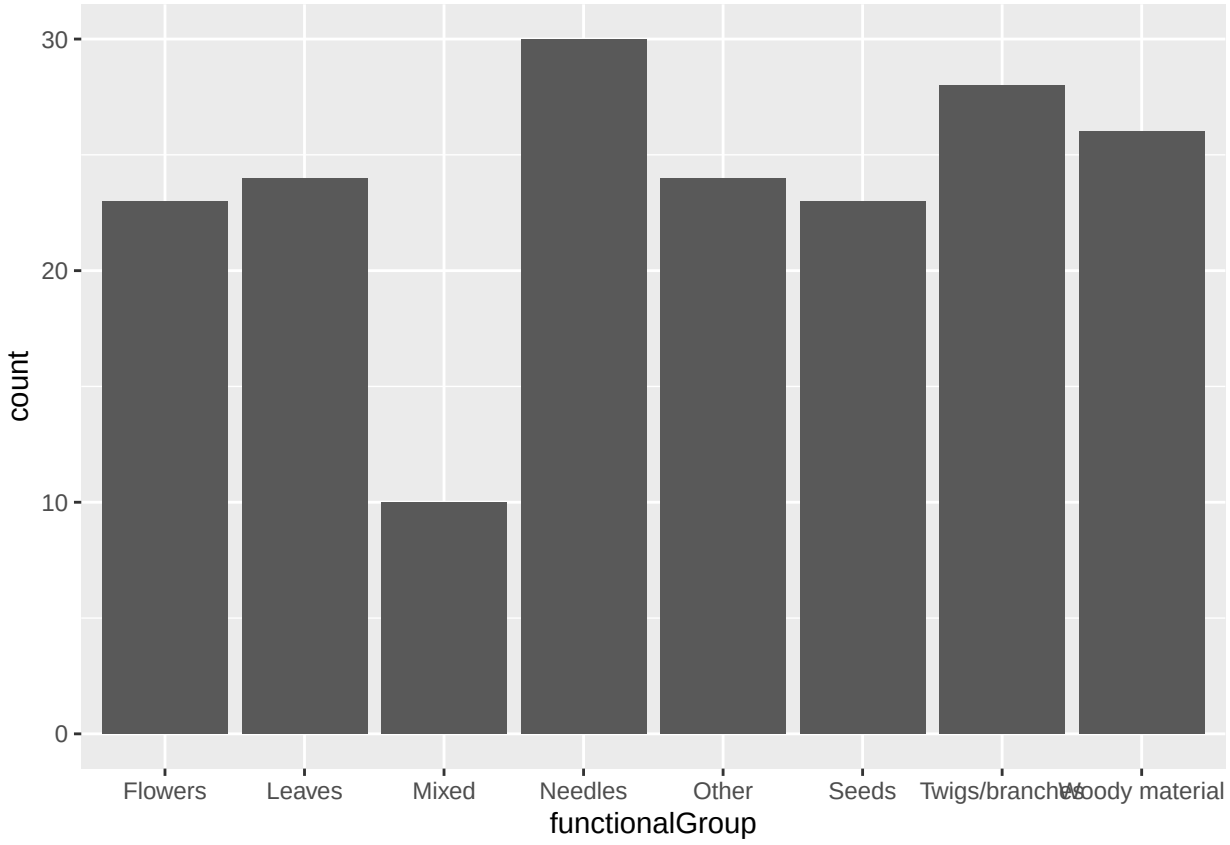
```
summary(Niwot_Ridge_NEON$plotID)
```

```
## 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 There were 12 plots sampled. The difference between the unique and summary function is that the summary shows the number of samples per plot, while the unqiue shows the number of plots only.

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.

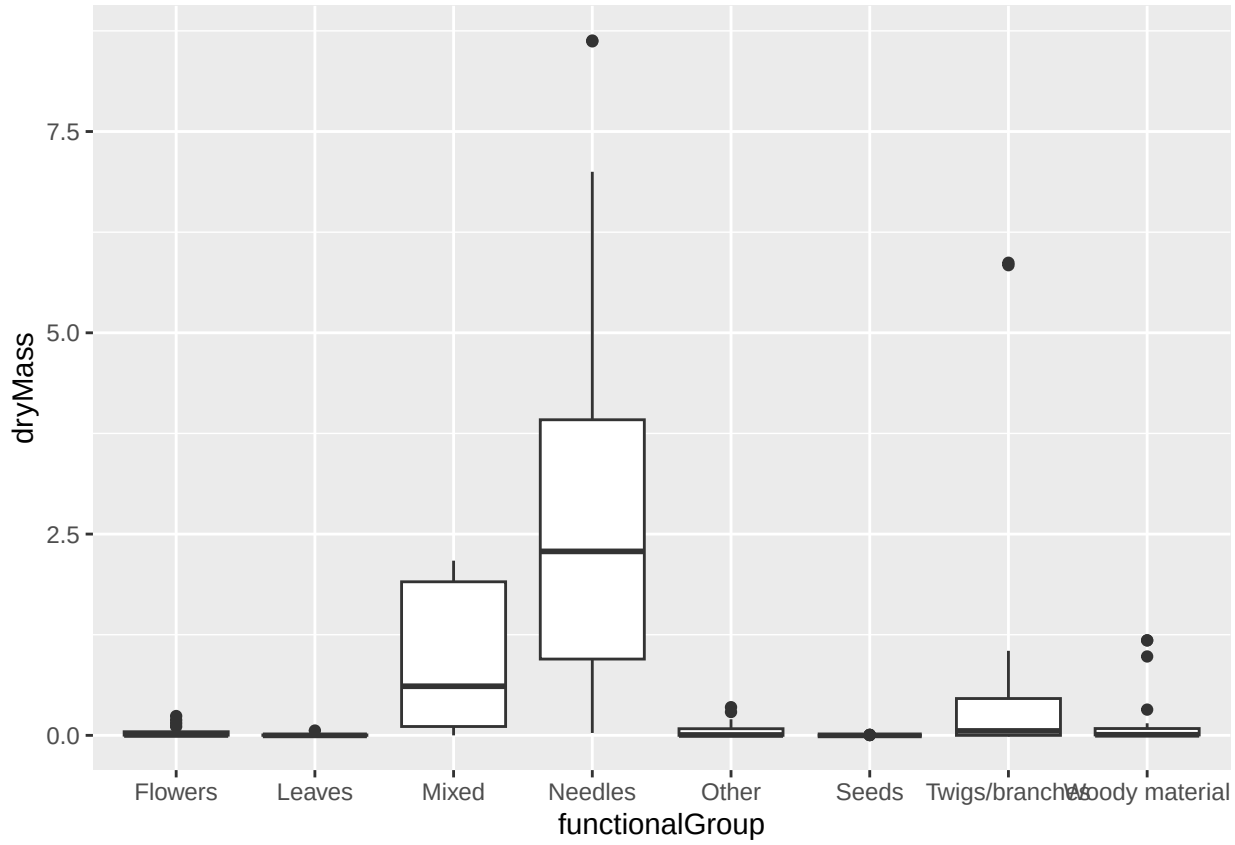
```
ggplot(Niwot_Ridge_NEON, aes(x = functionalGroup)) +
  geom_bar()
```



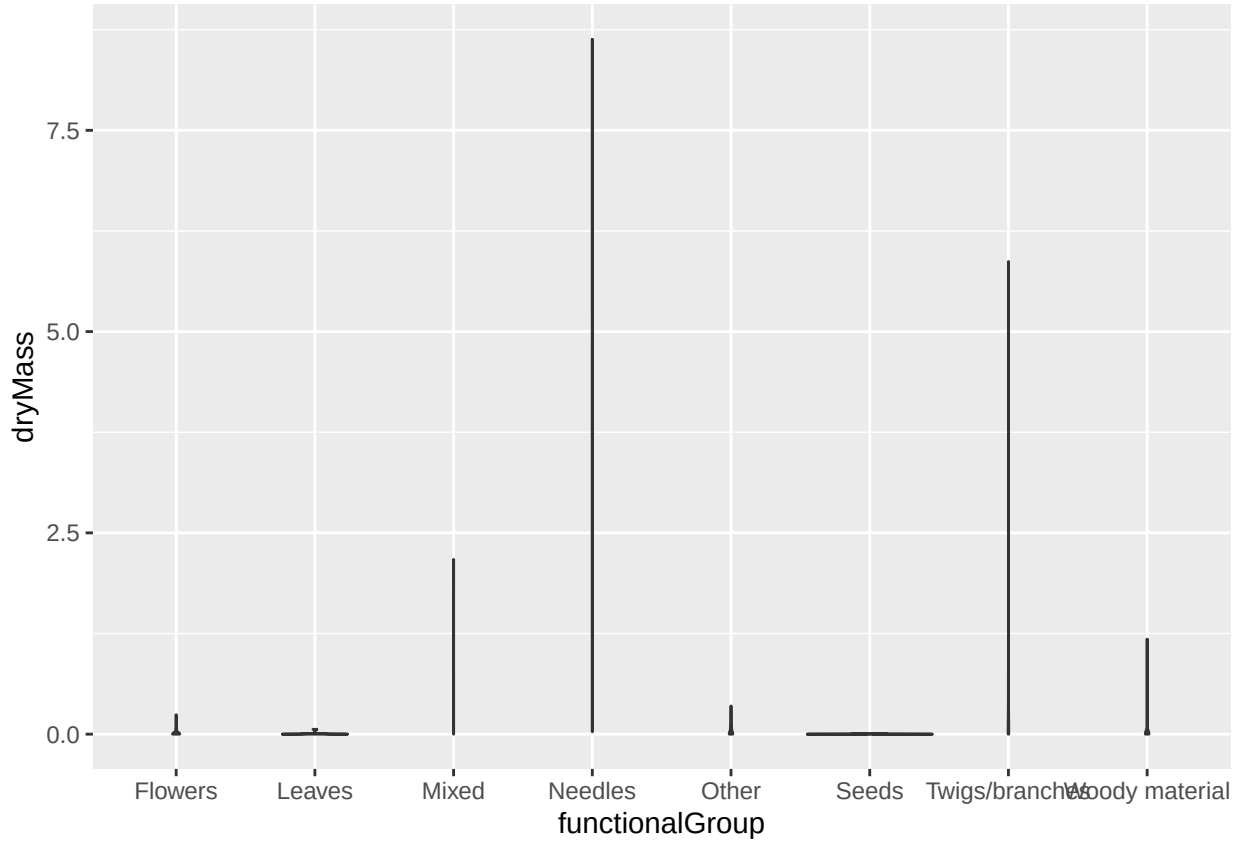
15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#boxplot

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




```
#violin plot
ggplot(Niwot_Ridge_NEON) +
  geom_violin(aes(x = functionalGroup, y = dryMass))
```



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

> Answer:The boxplot is more effective because we can see the distribution of values more clearly than in the violin plot.

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

> Answer:Needles have the highest biomass at these sites.