Assignment 3: Data Exploration

Lauren Shohan

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.

```
library(tidyverse)
library(lubridate)
library(here)
getwd()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```
here()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```
#ECOTOX neonicotinoid dataset
NeonicsData <- read.csv(
  file = here("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"),
  stringsAsFactors = TRUE)

#Niwot Ridge NEON dataset
LitterData <- 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 would be very interested to see how neonicotinoids impact insects on the area to see if this insecticide is killing or mutuating insects in different ways. This insecticide could be destroying crucial insect populations in the ecosystem that could damage the ecosystem and other species long term.

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 provides important insight on forest health and this woody debris supports biodiversity during decomposition and is essential to understand for forest management and conservation.

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 fine woody debris are collected from elevated and ground traps 2. Mass data for each collection event are measured into separate fuctional groups to an accuracy of 0.01 grams 3. Ground traps are sampled once a year while elevated trap samplings vary by vegetation on the site so it can be as short as 1-2 weeks or up to 2 months depending on the vegetation

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(NeonicsData)
```

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

```
#dimensions are 4623 rows by 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...]

```
sort(summary(NeonicsData$Effect)) #sorting Effect column
```

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

Answer: The top two most common effects studied are mortality (1493) and population (1803). These could be of specific interest to know how a population is doing, if they are thriving/dominating an environment with a higher population rate than mortality or reverse and if their death rate is much higher than population and find the factors that are influencing this decline or increase.

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(NeonicsData$Species.Common.Name, maxsum = 7)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee      Italian Honeybee
##                   152                   140                   113
##               (Other)
##                  3083
```

```
#did 7 so that I could see the top 6 species not including the Other category
```

Answer: The top 6 most commonly studied species (not including the other category), are Honey Bee, Parasitic wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. All these species are flying pollinators and thus it would be very important and interesting to study these species since they play such an important role in our ecosystems.

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(NeonicsData$Conc.1..Author.)
```

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

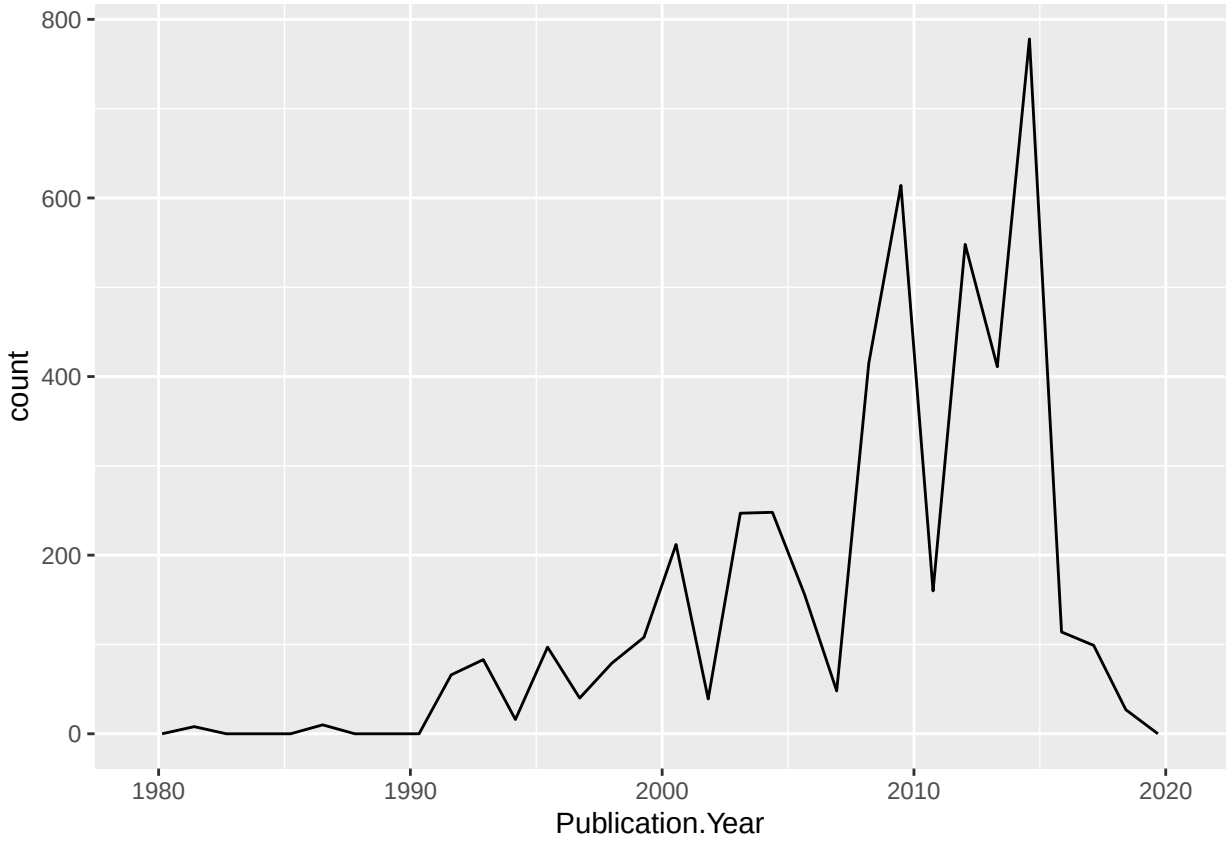
Answer: The class of this column is a factor because the column is full of different labels like "Active Ingredient" and "Formulation" and numeric cannot handle string types, only numbers.

Explore your data graphically (Neonics)

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

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

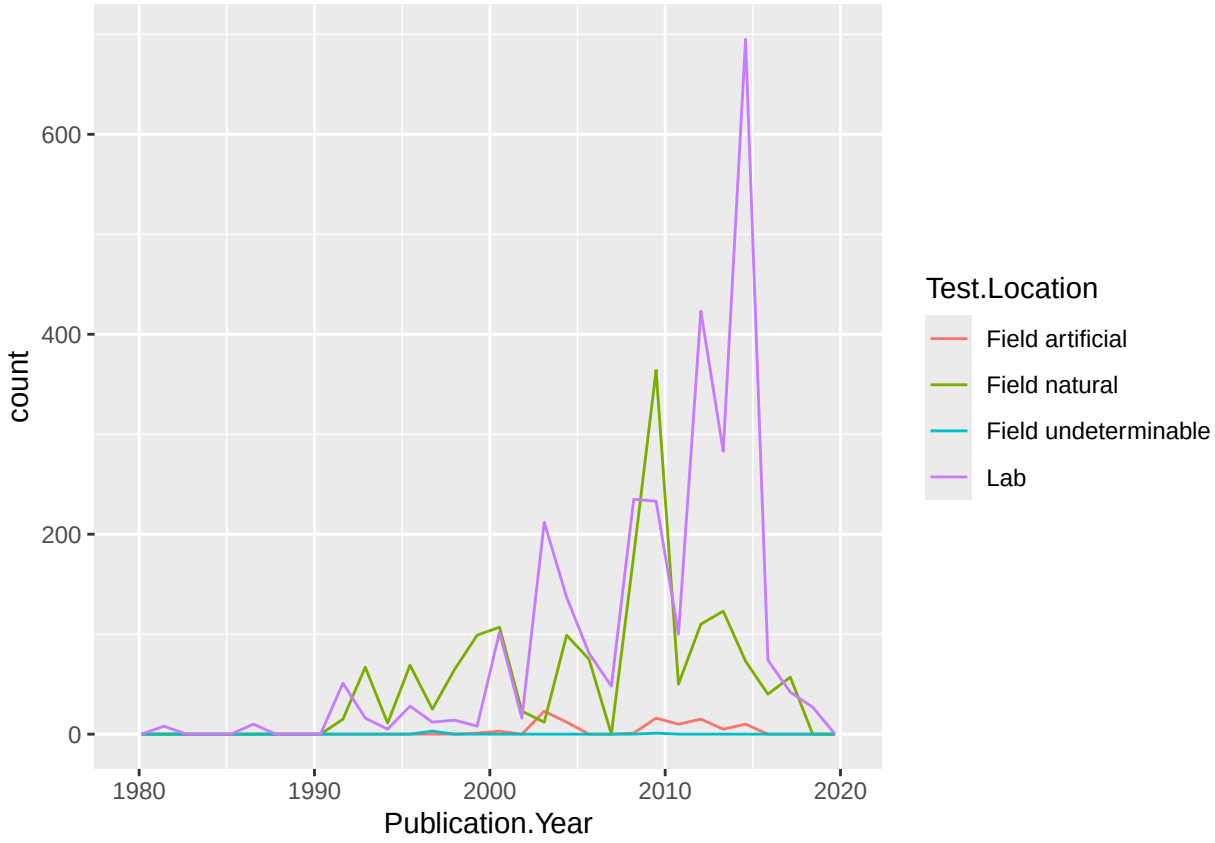
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
ggplot(NeonicsData) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location))
```

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




```
#added test.location as color
```

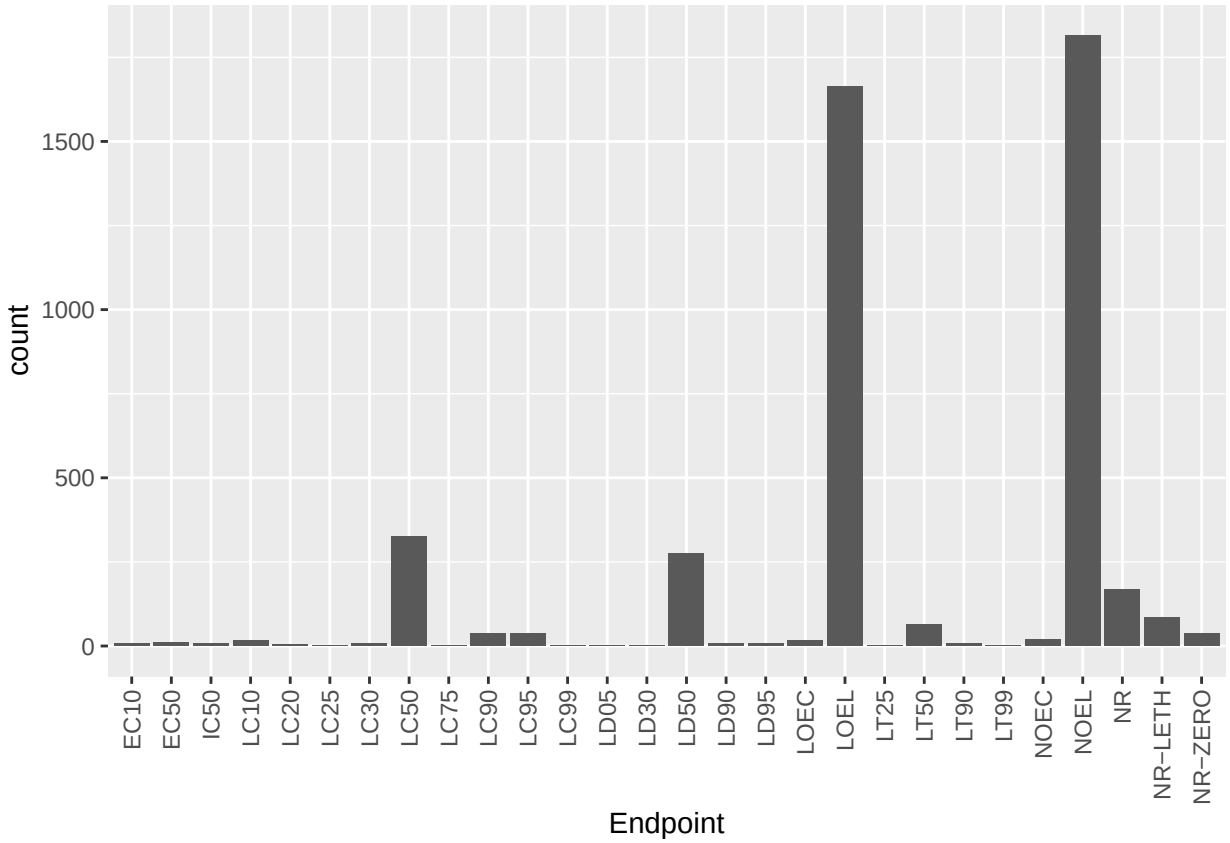
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The lab is definitely the most common test location at least from 2000 until now. Before 2000, the natural field was the most common, but the lab skyrocketed around 2003, but there was one final spike of natural field in 2009 before lab locations dominated especially around 2013.

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(NeonicsData) + #importing data
  geom_bar(aes(x = Endpoint)) + #setting my x
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: The two most common points are LOEL and NOEL. LOEL stands for Lowest-observable-effect-level while NOEL stands for No-observable-effect-level and they are both from the terrestrial database.

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(LitterData$collectDate) #it was a factor
```

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

```
LitterData$collectDate <- as.Date(LitterData$collectDate)
#assigning that column as a date and then reassigning into overall dataframe - it is now a date
class(LitterData$collectDate)
```

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

```
unique(LitterData$collectDate)
```

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

```
  #2018-08-02 and 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(LitterData$plotID)
```

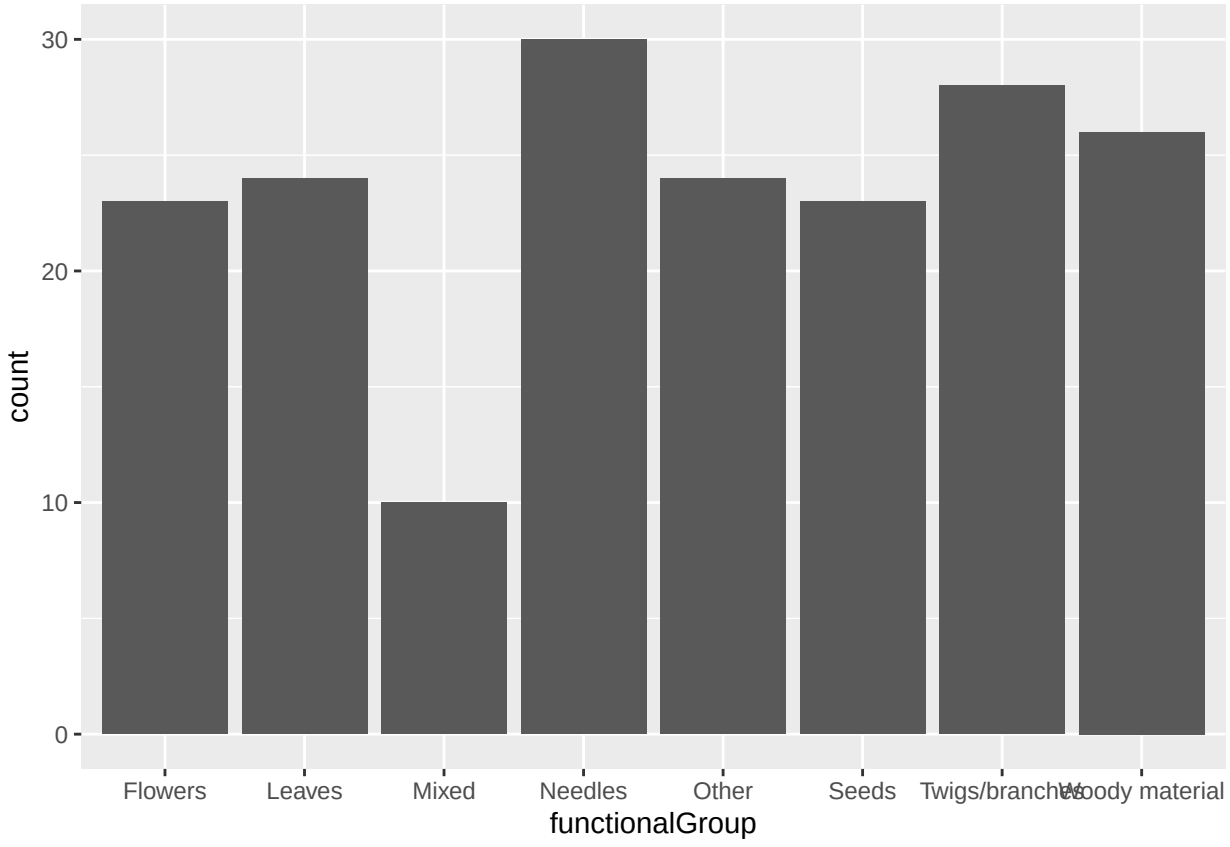
```
##  [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

```
#summary(LitterData$plotID)
```

Answer: Using the unique function, I found that 12 plots were sampled at Niwot Ridge. This information obtained from unique() returns the disctinct values in the dataset and removes duplicates while summary() summarizes all the information and gives a broader overview of the data. Summary gave me the counts for all the different plots while unique told me there were 12 unique plot Ids.

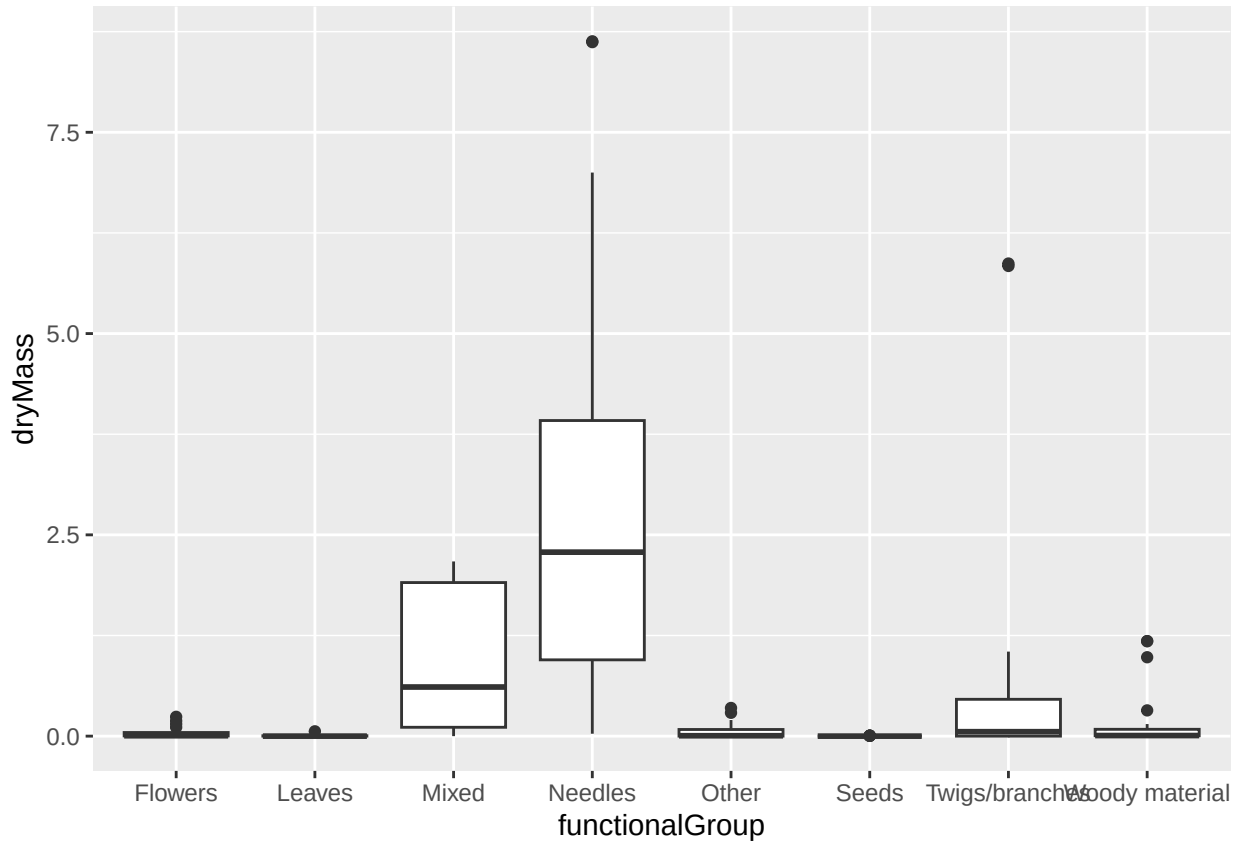
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(LitterData) +
  geom_bar(aes(x = functionalGroup)) #assigning functiongroup as my x
```

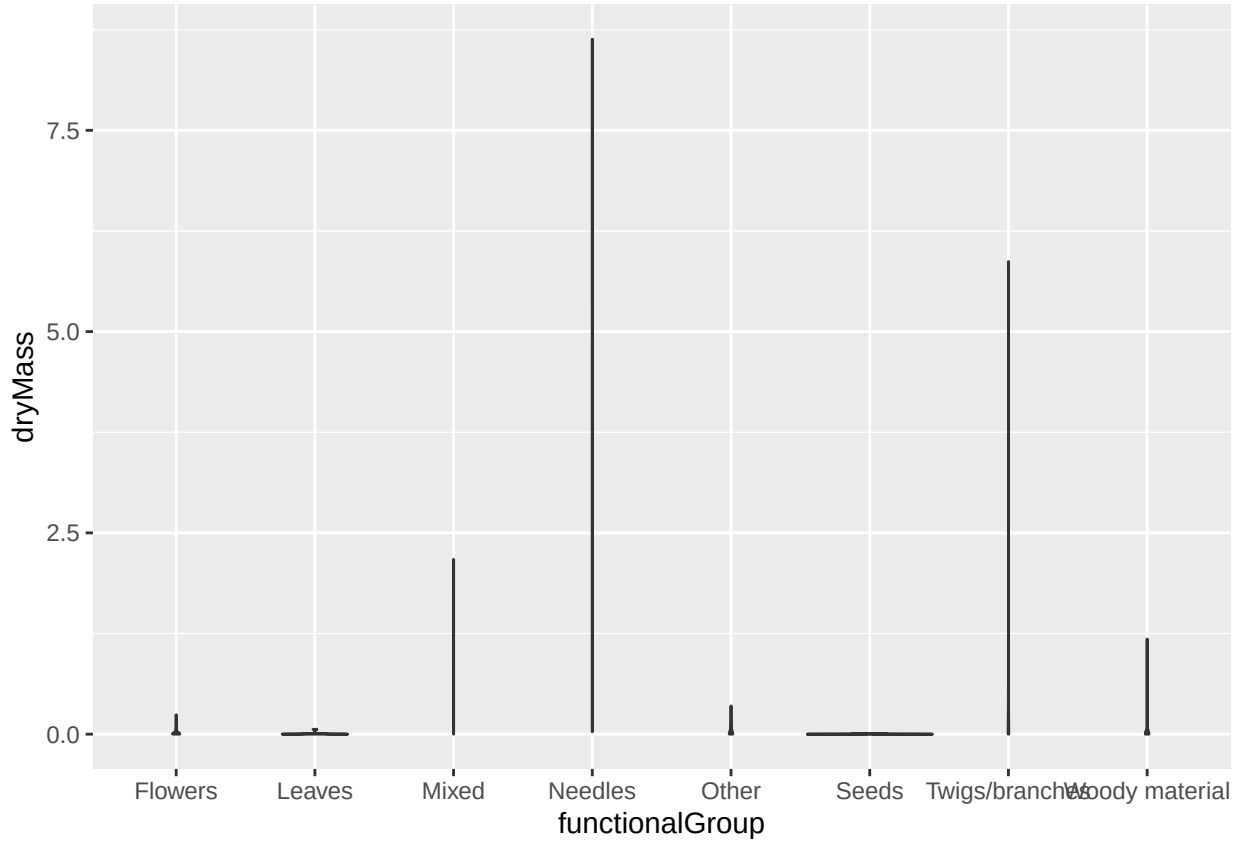


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(LitterData) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass)) #did my continuous data as my x
```




```
ggplot(LitterData) +
  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 blox plot is much easier to visually understand because the data in violin plot is very clustered around really low masses. This cluttered data makes it very difficult to interpret the differences in groups.

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

> Answer: Needles have the highest biomass at these sites with a mean of just under 2.5, higher than all the other ones. Mixed and twigs/branches come in second and third.