Assignment 3: Data Exploration

Meg Manning

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.

```
#check working directory - in EDE_Fall2023 folder
getwd()
```

```
## [1] "/Users/margaretmanning/Desktop/GitHub/ENVIRON 872/EDE_Fall2023"
```

```
#load packages (tidyverse, lubridate)
library(tidyverse)
library(lubridate)

#Upload 2 datasets - ECOTOX and Niwot Ridge NEON with stringsAsFactors = TRUE
```

```
Neonics <- read.csv("/Users/margaretmanning/Desktop/GitHub/ENVIRON 872/EDE_Fall2023/Data/Raw/ECOTOX_Neon
Litter <- read.csv("/Users/margaretmanning/Desktop/GitHub/ENVIRON 872/EDE_Fall2023/Data/Raw/NEON_NIWO_Li

View(Neonics)
View(Litter)
```

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: When looking up more about neonicotinoids, it says that they can attack the nervous systems of insects, including bees in particular. Bees are an incredibly important insect for pollination and act as a keystone species in sustaining ecosystems. This being said, an insecticide, like neonicotinoid, is killing off bees and as a result, hurting the wider ecosystem. Studying the ecotoxicology of neonicotinoids can help mitigate and regulate the harmful effects this insecticide has on bees.

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 can act as a nursery ground for other organisms to grow. One example is a nursery log that you can typically find in forests in the Pacific Northwest. Trees that fall in old growth forests provide nutrients and space for other plants and fungi to grow. Litter and woody debris are key parts of forest ecosystems. In this case, they can also give insight into forest decay rates.

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. Fine wood debris is collected in ground traps 2. Litter is collected in elevated PVC litter traps have a mesh "basket" located ~80cm from ground 3. Sampling only occurs in tower plots which are randomly selected and trap placement within plots may either be targeted or randomized, depending on the vegetation

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#use dim() to see the dimensions of both datasets
dim(Litter)
```

```
## [1] 188  19
```

```r
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?

```r
#use summary paired with a $ to point to the Effect column to see the summary stats for this column
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: These effects could be specifically of interest beause it is giving important information about the overall sample mortality, physiology, reproduction, etc., which are important metrics for better understanding the effect of neonicotinoid insecticides on the types of insects they're studying.

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 summary function $Species.Common.Name, 6 to see the top 6 results for most commons species
summary(Neonics$Species.Common.Name, 6)
```

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

Answer: The top 6 most common species in order are Honey Bee, Parasitic Wasp, Tailed Bumblebee, Carnionlan Honey Bee, and Bumble Bee. All of these species are types of bees. This is of interest over other insects because bees are keystone species for ecosystem management and sustainability.

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

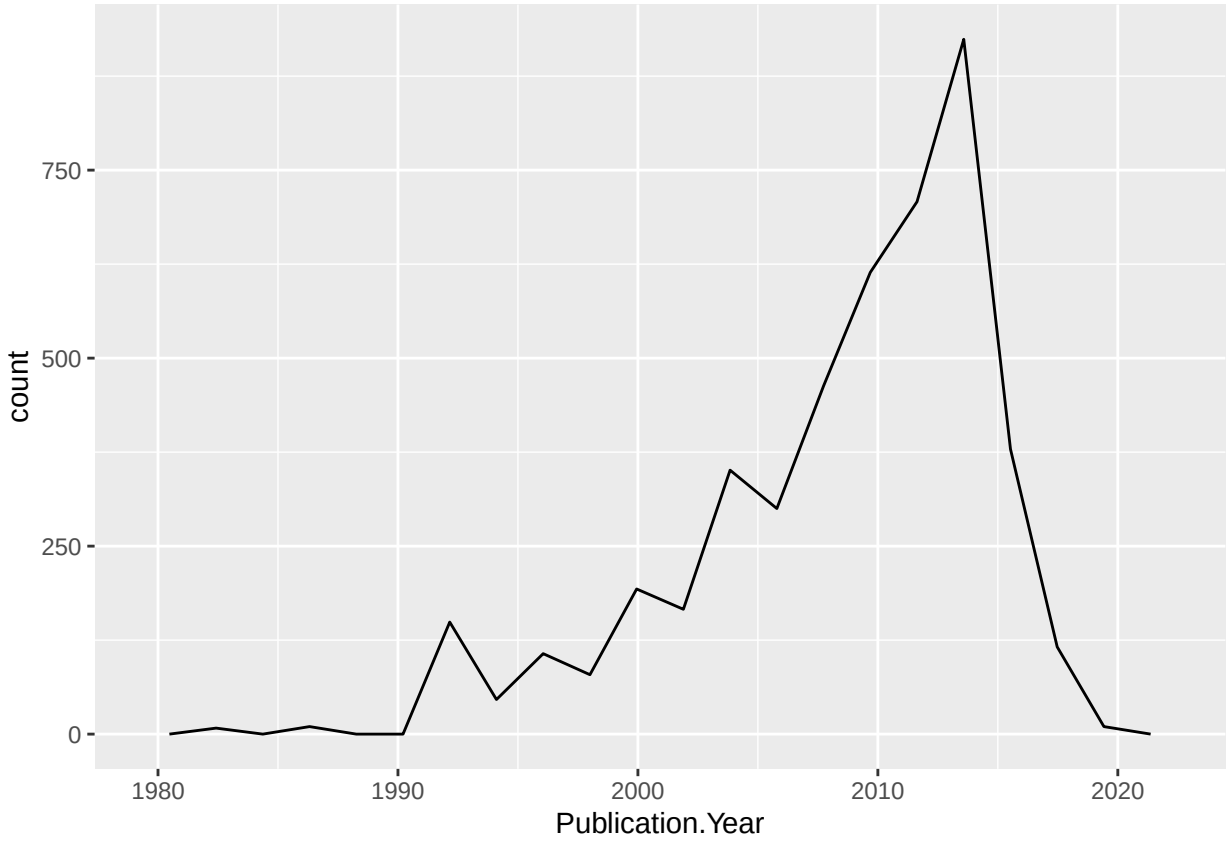
```
## [1] "factor"
```

Answer: It is not numeric because some of the values in that column are not numeric values.

Explore your data graphically (Neonics)

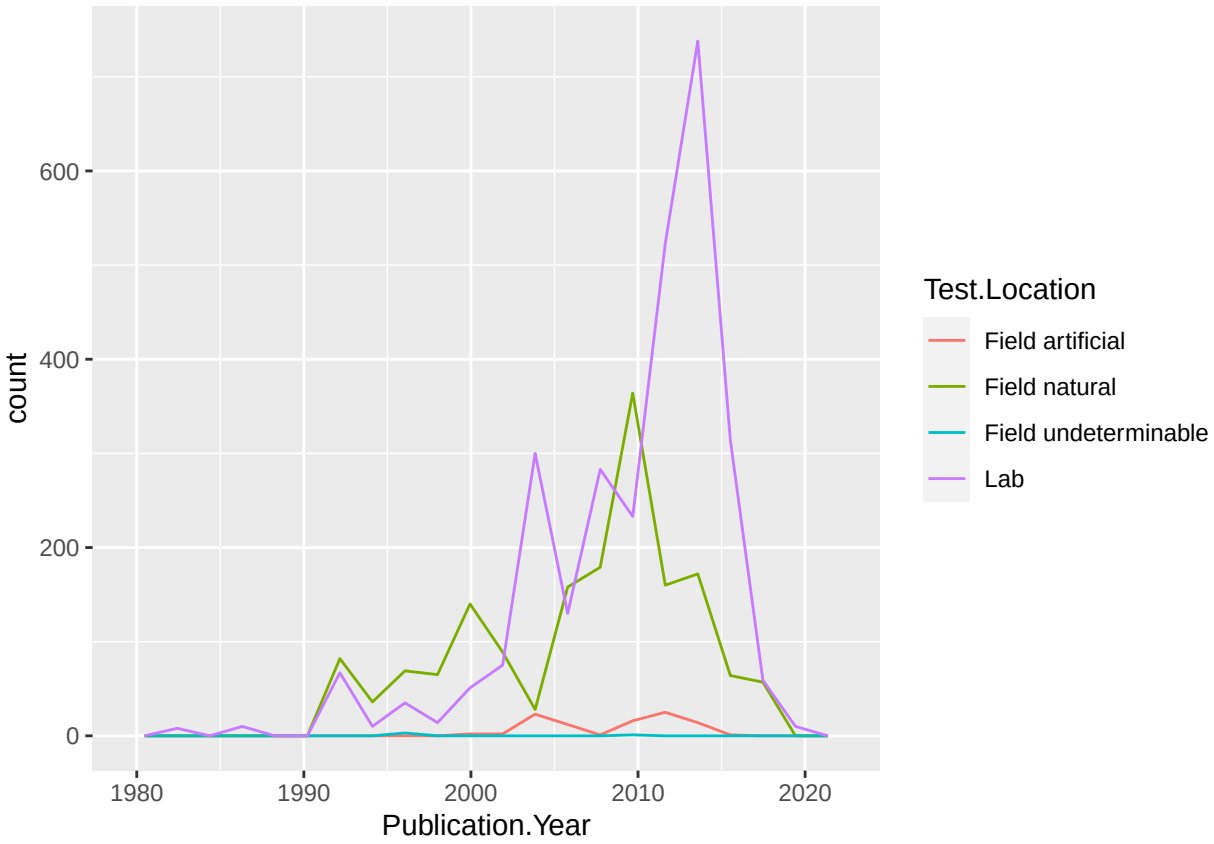
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
#create a plot of the number of studies conducted by publication year
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 20)
```



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

```
#add a color aesthetic for each Test.Location (change the aesthetics)
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 20)
```



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

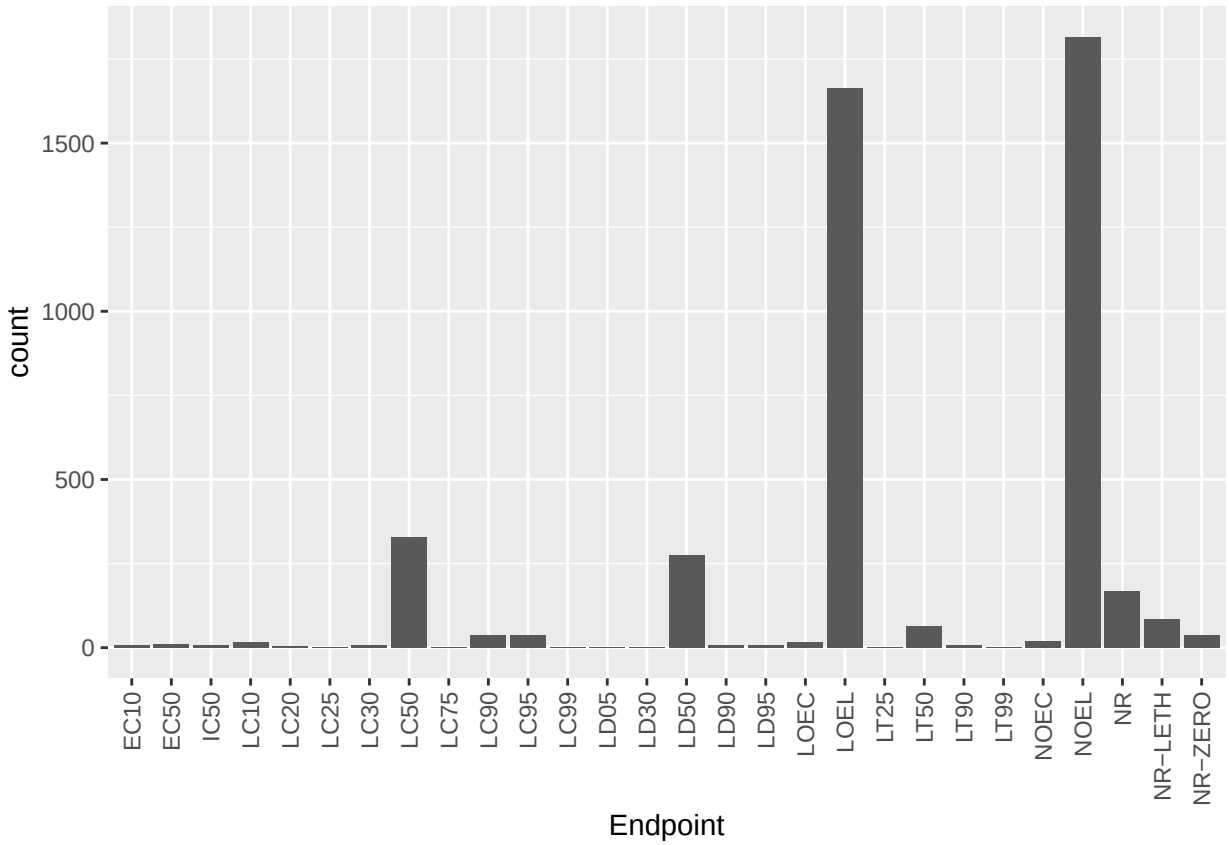
> Answer: The most common test locations are in a Lab. As shown in the graph, the overall count of Lab test locations appears to be much higher than the other three options. The second most common locations seem to be Field Natural. Over time both Lab and Field Natural locations seem to go up in the early 2000s and then decline steeply around either 2010 or just after. Around 2020 the number of location sites for all four places seems to be around the same as in 1990.

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...]

```
#create a bar graph
ggplot(Neonics) +
  geom_bar(aes(x = Endpoint), bins = 20) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

```
## Warning in geom_bar(aes(x = Endpoint), bins = 20): Ignoring unknown parameters:
## `bins`
```



Answer: The two most common endpoints are LOEL and NOEL. In terms of the chemical grade and standards code the -EL means Electrophoresis Grade which is defined as material used specifically for electrophoresis applications.

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.

```
#check the class of the date
class(Litter$collectDate)
```

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

```
#convert from factor to date
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y%m%d")

#check saved as date
class(Litter$collectDate)
```

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

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`?

```
#find how many plots were sampled at Niwot Ridge
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
```

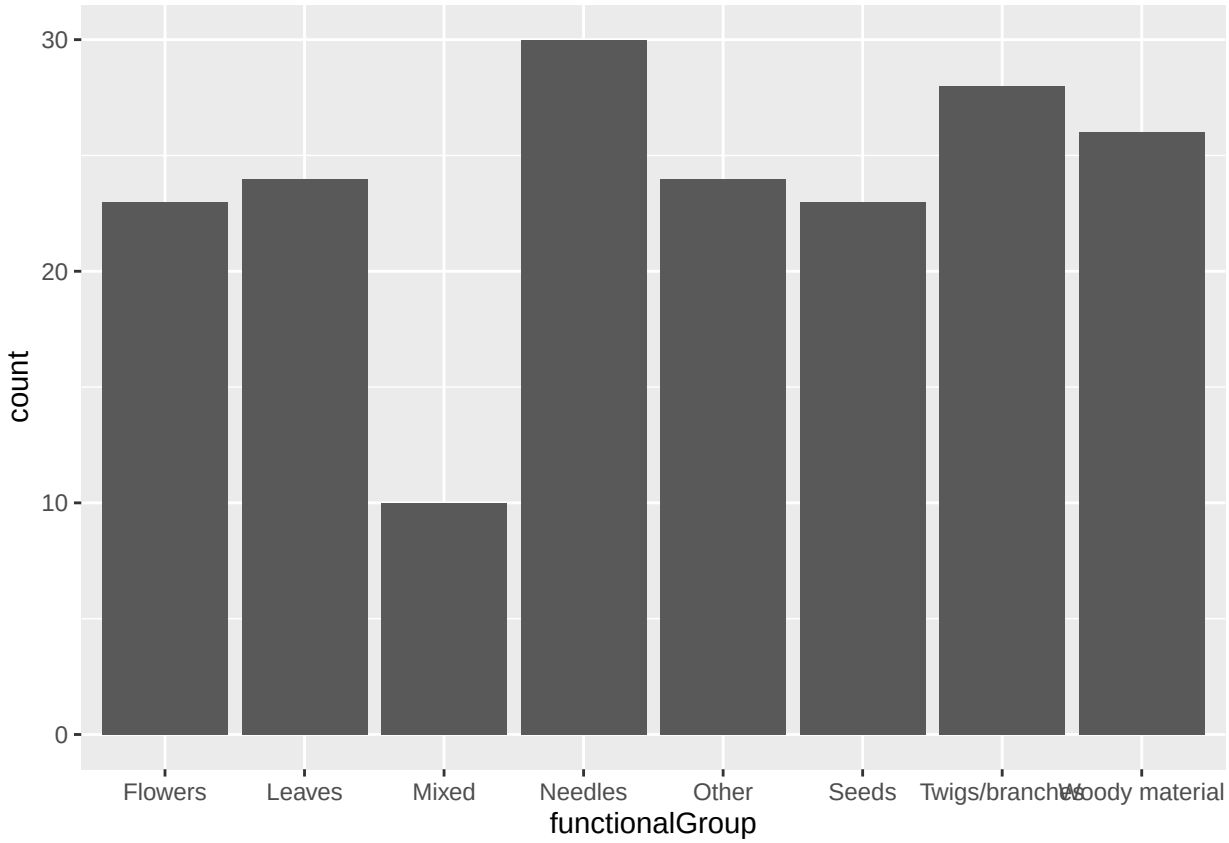
```
summary(Litter$plotID)
```

```
## 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 difference between the two functions is that the unique() function provides the number of sites at Niwot Ridge, but it does not tell you how many plots were sampled at each site at Niwot Ridge. The summary function provides each location and the number of samples at each site.

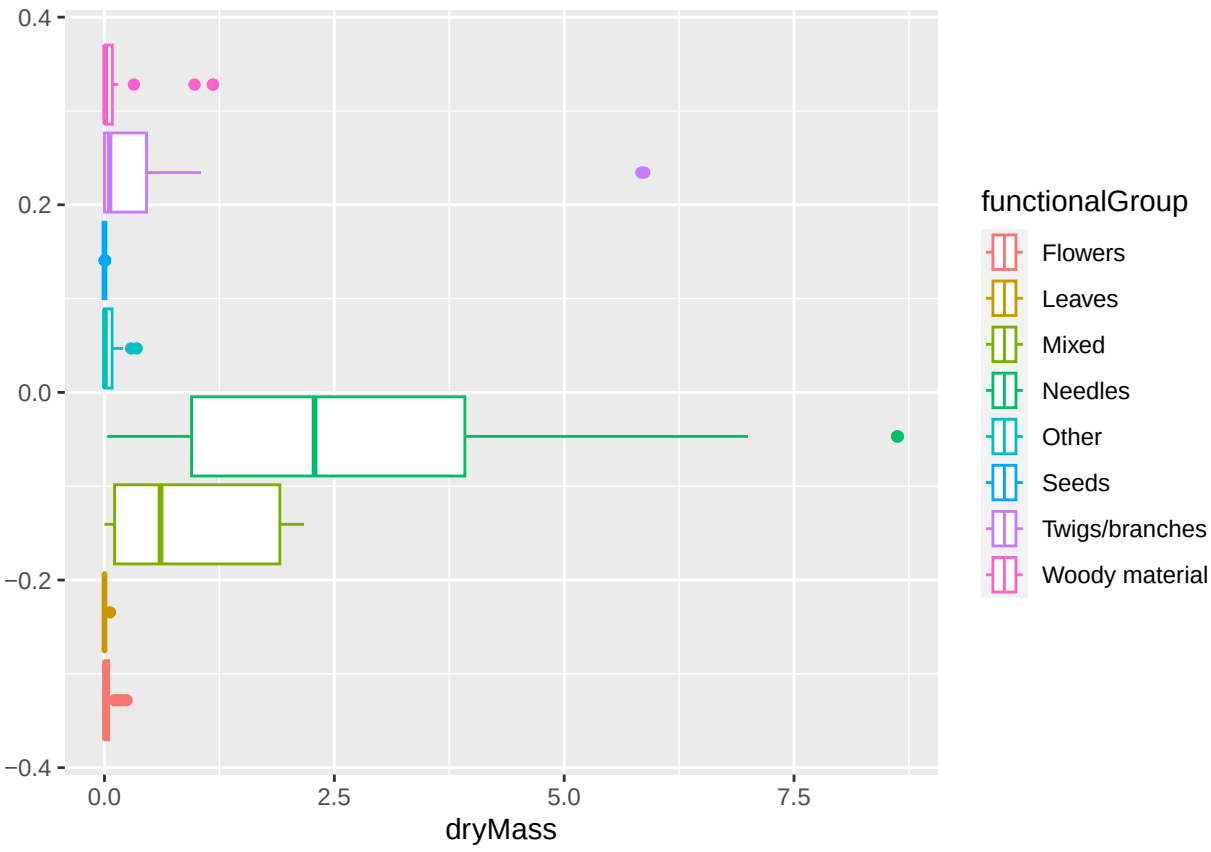
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#create a bar graph of functionalGroup
ggplot(Litter) +
  geom_bar(aes(x = functionalGroup))
```

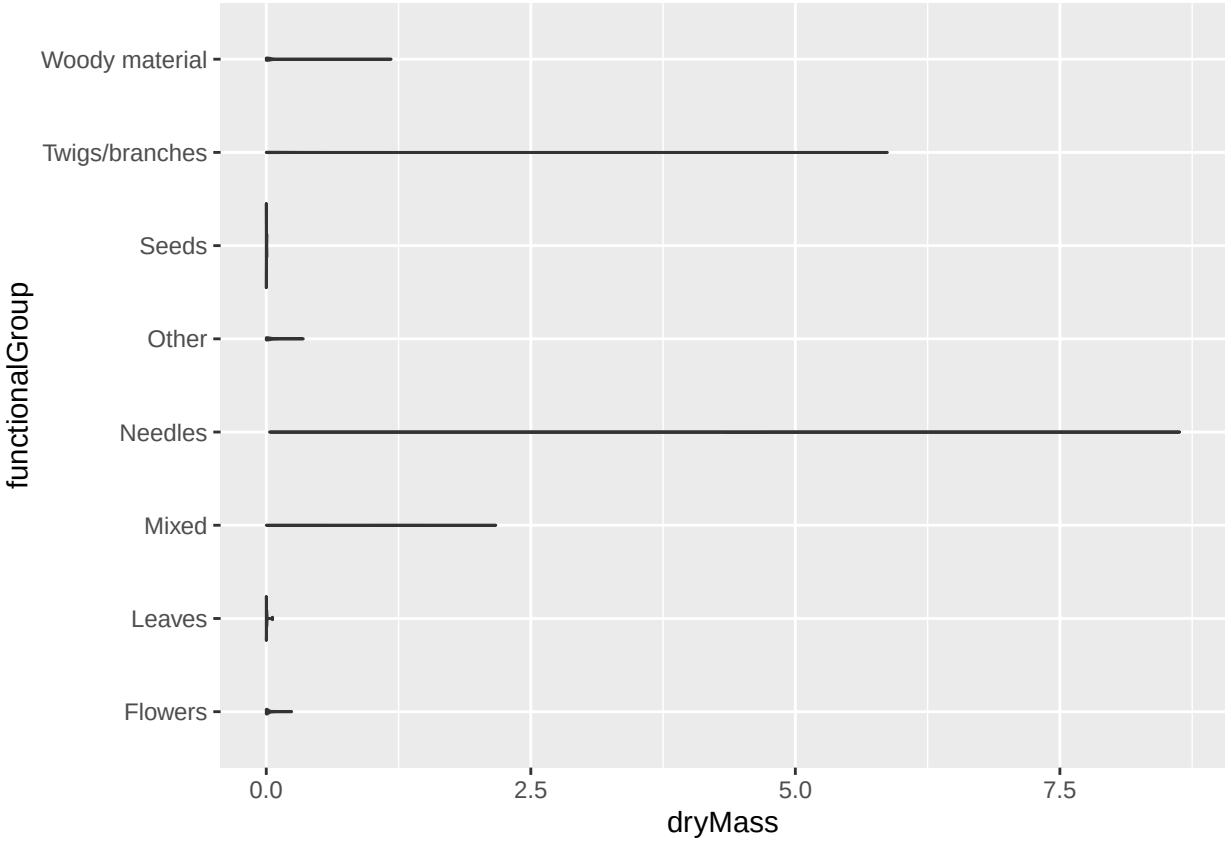


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#create a boxplot of dryMass by functionalGroup
ggplot(Litter) +
  geom_boxplot(aes(x = dryMass, color = functionalGroup))
```




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



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

> Answer: The box plot is better because it gives you the quick statistics for items in each functional group, allowing us to see things like median, IQR, and outliers. The violin plot looks at the distribution of the data, which is harder in this case to see what each functional group looks like without zooming in to each.

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

> Answer: The types of litter that have the highest biomass is by far needles. Using the boxplot as a reference, we can see that the median is significantly high that all of the other types of litter. The second most appears to be the "mixed" type. However, the twigs/branches have an outlier that is shown to be above the median of both the needles and mixed biomass types.