Assignment 3: Data Exploration

Meghan Seyler, Section #2

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Exploration.

Directions

1. Change "Student Name, Section #" on line 3 (above) with your name and section number.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "FirstLast_A03_DataExploration.Rmd") prior to submission.

The completed exercise is due on <>.

Set up your R session

1. Check your working directory, load necessary packages (tidyverse), 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 add the `stringsAsFactors = TRUE` parameter to the function when reading in the CSV files.**

```
getwd()
```

```
## [1] "Z:/ENV872/Environmental_Data_Analytics_2022/Assignments"
```

```
library(tidyverse)

Neonics<-read.csv("../Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",
                  stringsAsFactors = TRUE)

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 ecotoxicologoy of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: The ecotoxicology of neonicotinoids is relevant because it is important to understand how a widely used inescticied affects different species of insects and animals. From a brief look at the data, I see that this data set also recorded dosage. So one thing we may be interested in looking into is the minimum dosage required to have the intended affect on a particular insect.

This way we can minimize the use of insecticides and avoid any potential consequences of using a higher than needed dose.

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 measurements can tell us about Aboveground Net Primary Productivtiy.This information looks like it was collected at many different sites and around the same date.

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: *Litter is defined as material that is dropped from the forest canopy and has a butt end diameter of <2cm and a length <50cm* Litter is collected from elevated traps and fine woody debris is collected from groud traps. *All reports are of a single trap at a single point in time.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Litter) # 188 rows and 19 columns
```

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

```
dim(Neonics) #4623 rows and 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?

```
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: It is useful to get this quick view of frequency of effects of the insecticide because it allows the analyzer to start forming more questions. For example I see that population effects and mortality effects are by far the most frequent, based on this information I may want to investigate the doeses used that caused these effects. Additionally, it is interesting that only one species expeprienced hormonal effects. Overall it is a good place to start understanding the story of the data.

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.

```
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: They're all in the bee category indicating they are probably gentically similar. This indicates that the insecticide is clearly effecting a specific species more than others. Additionally, it is good to know that the "other" category occupies such a high number of samples.

8. Concentrations are always a numeric value. What is the class of Conc.1..Author. in the dataset, and why is it not numeric?

```
typeof(Neonics$Conc.1..Author.)
```

```
## [1] "integer"
```

```
class(Neonics$Conc.1..Author.)
```

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

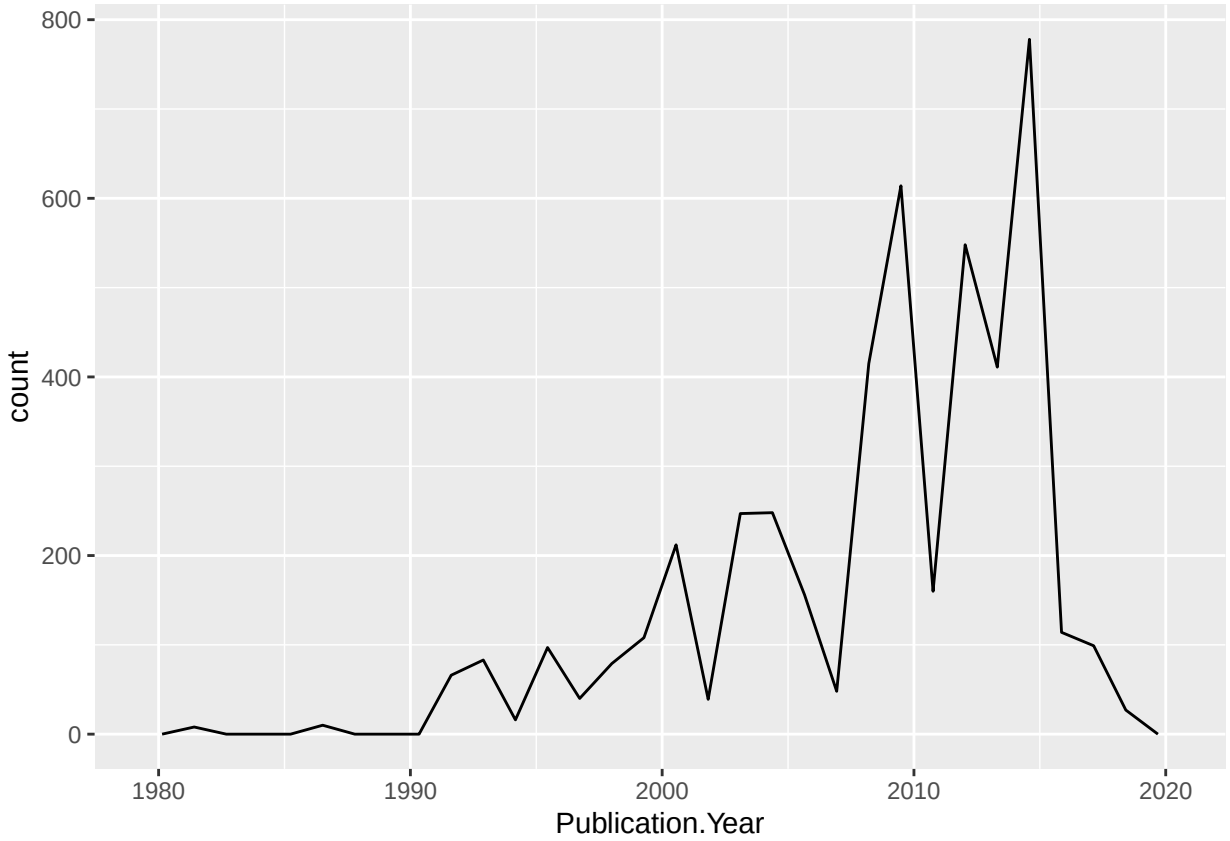
Answer: Conc.1..Author is classified as a factor. It is not numveric because there is a categorical element to this column.

Explore your data graphically (Neonics)

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

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

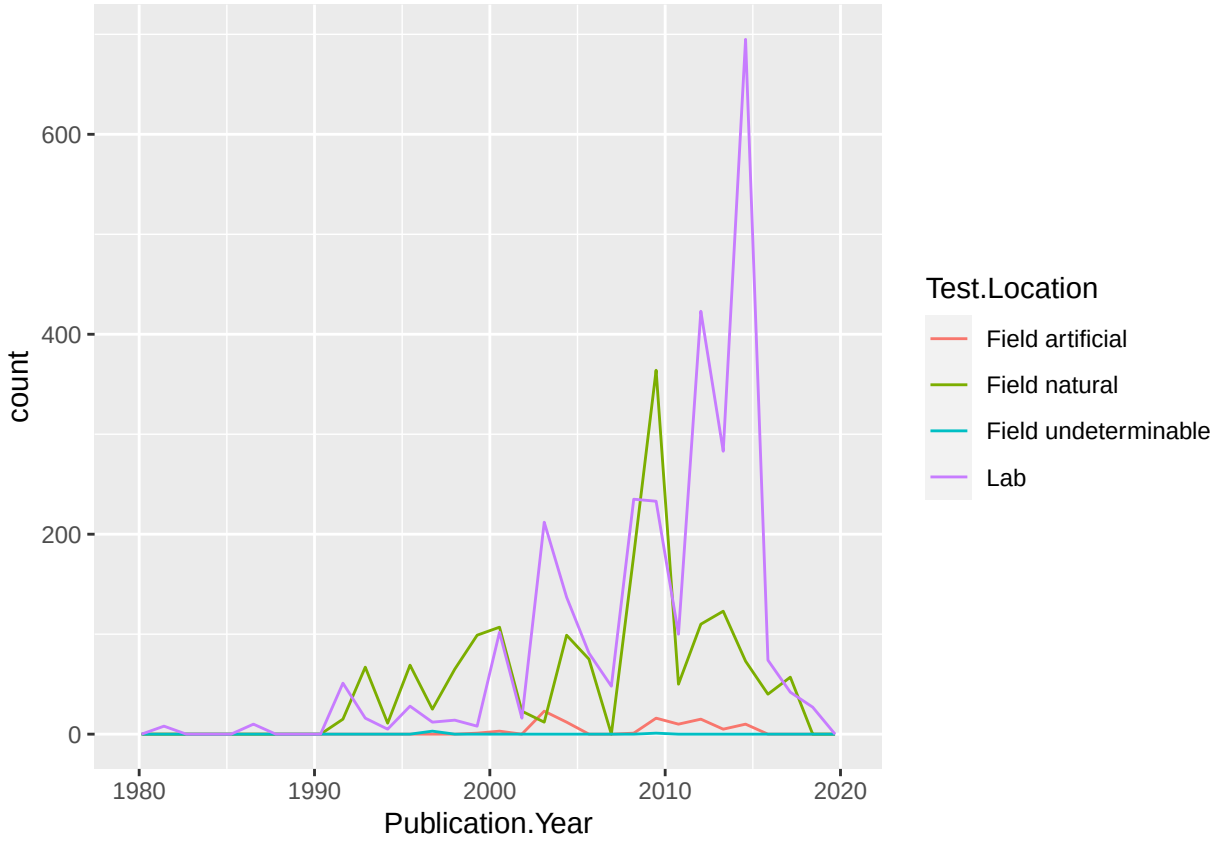
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

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

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

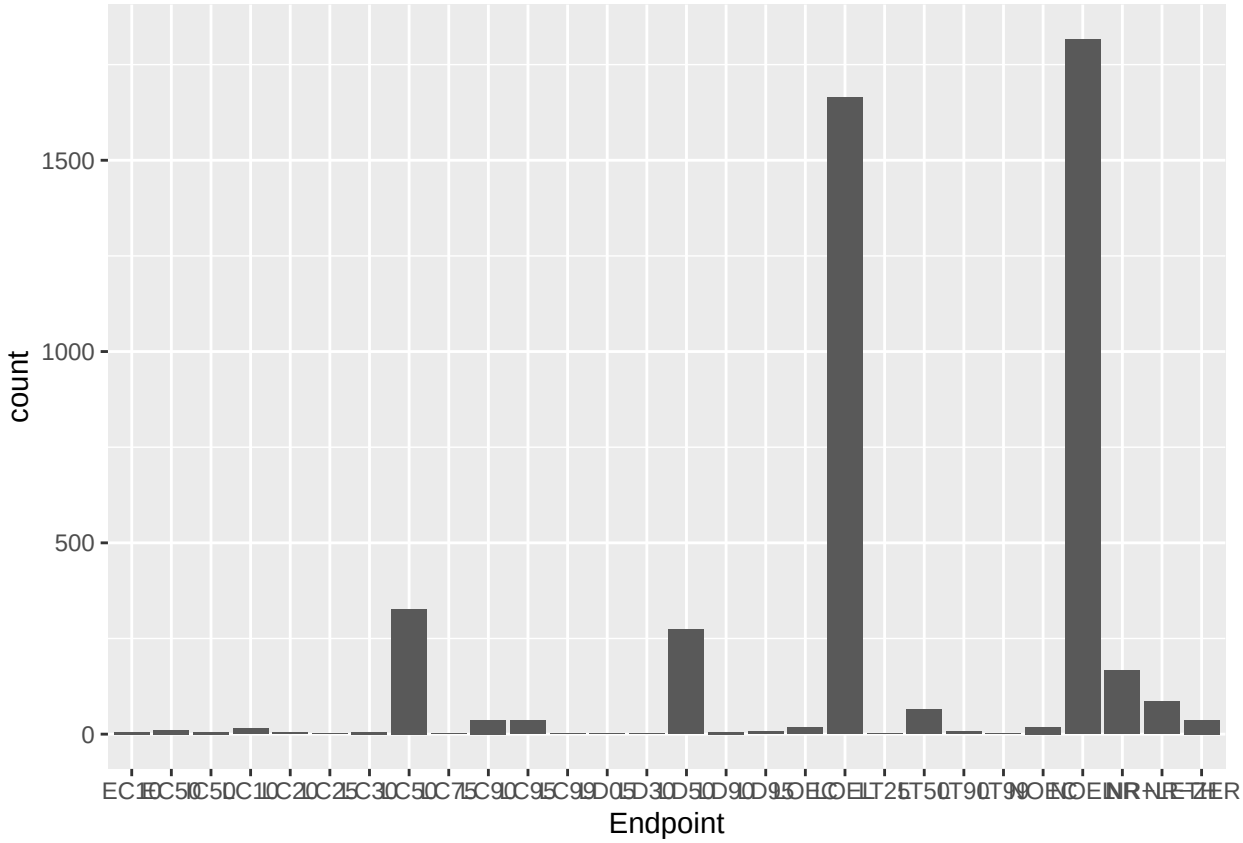


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are Lab and Field natural. While these are consistently the two most common test locations, they do change a bit overtime. For example Field natural was the most common from about 1992-2000 but from 2000-2020 Lab was consistently most common except for in 2009.

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.

```
ggplot(Neonics)+
  geom_bar(aes(x=Endpoint))
```



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

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) #collectDate class = factor
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")

class(Litter$collectDate) #checking that class changed to Date
```

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

```
unique(Litter$collectDate) #liter was sampled on August 2nd and 8th 2018
```

```
## [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(Litter$plotID) #12 unique plots were sampled at Niwot Ridge
```

```
##  [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
```

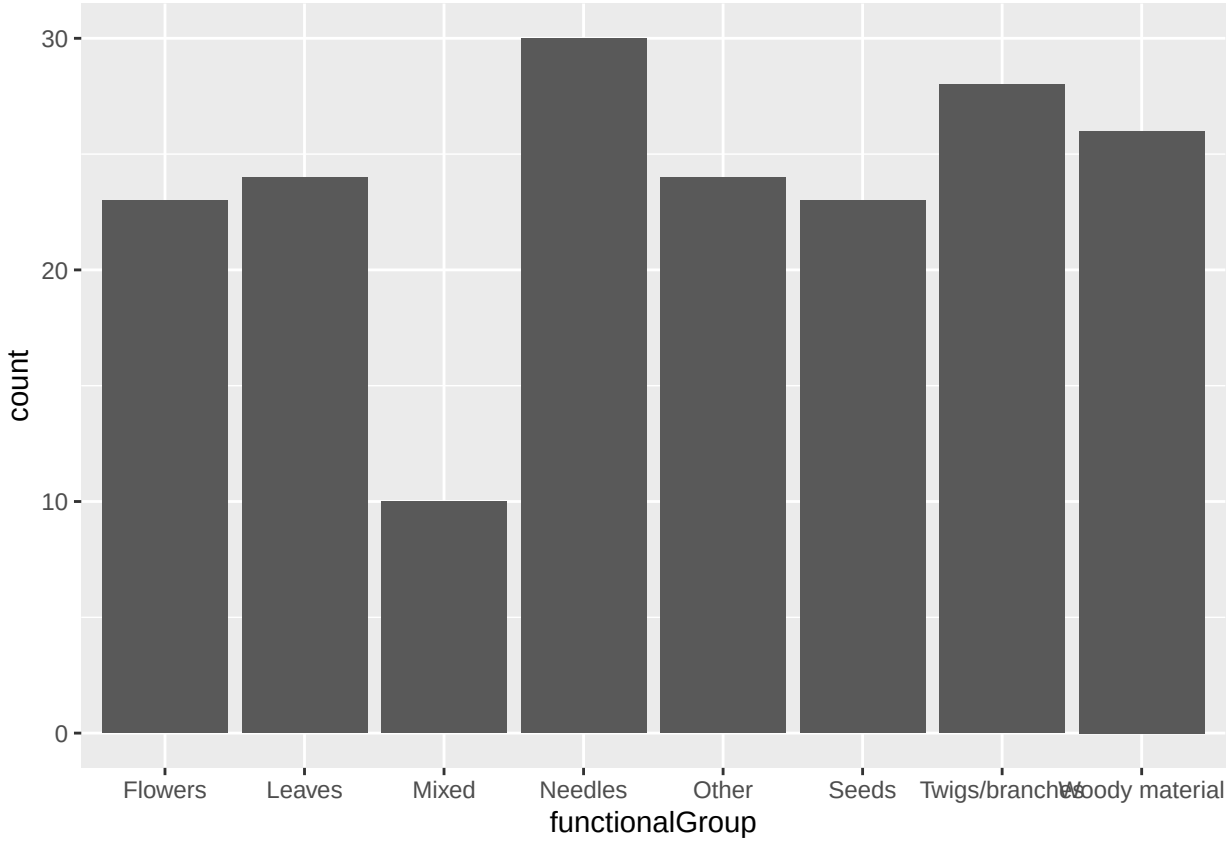
```
summary(Litter$plotID)
```

```
## 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 unique function resulted in an output that told us the number of unique plots sampled. The summary funciton is different than unique because it showed each of the unique plots as well as the number of plots within each unique category. Also, the summary output does not ouput a number communicating how many total unique plots there are so if this were a larger data set it would be best to use both of these commands together to best understand the data.

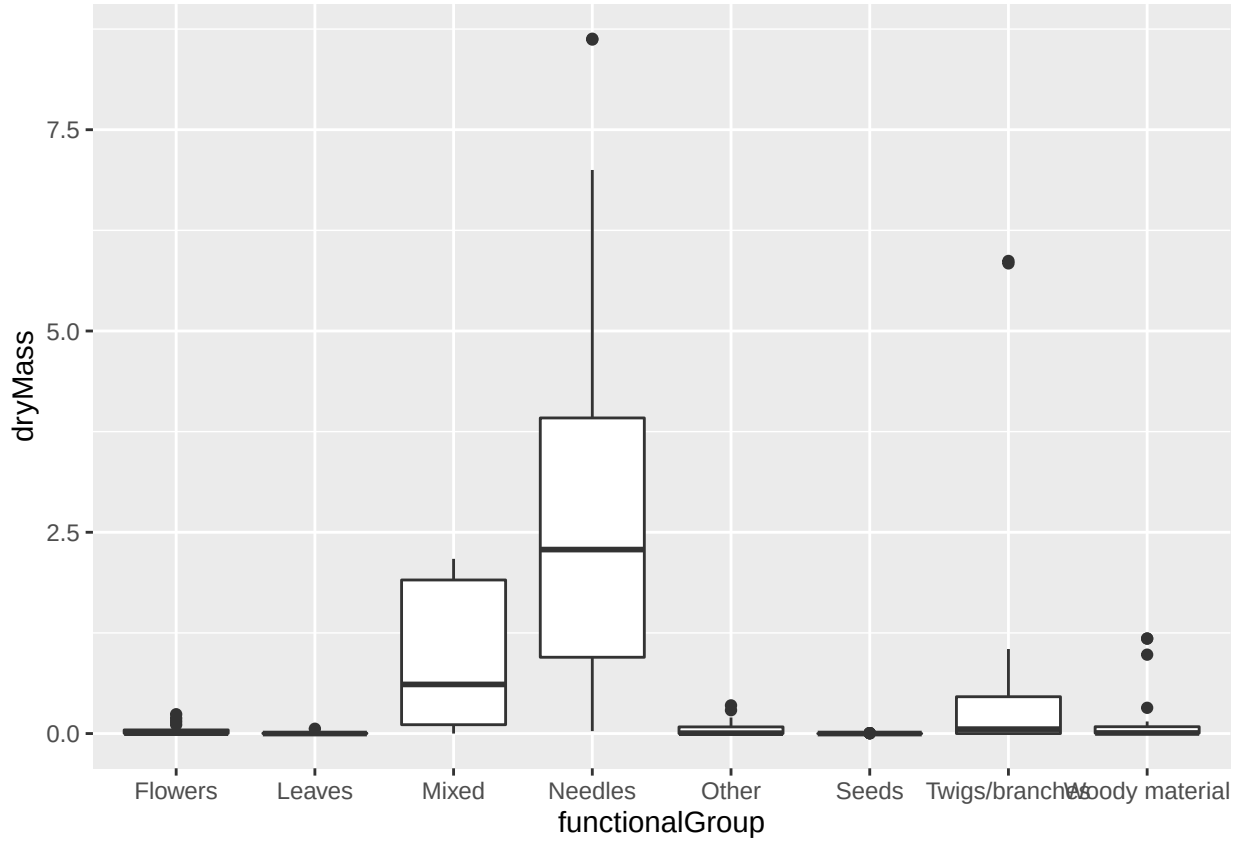
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

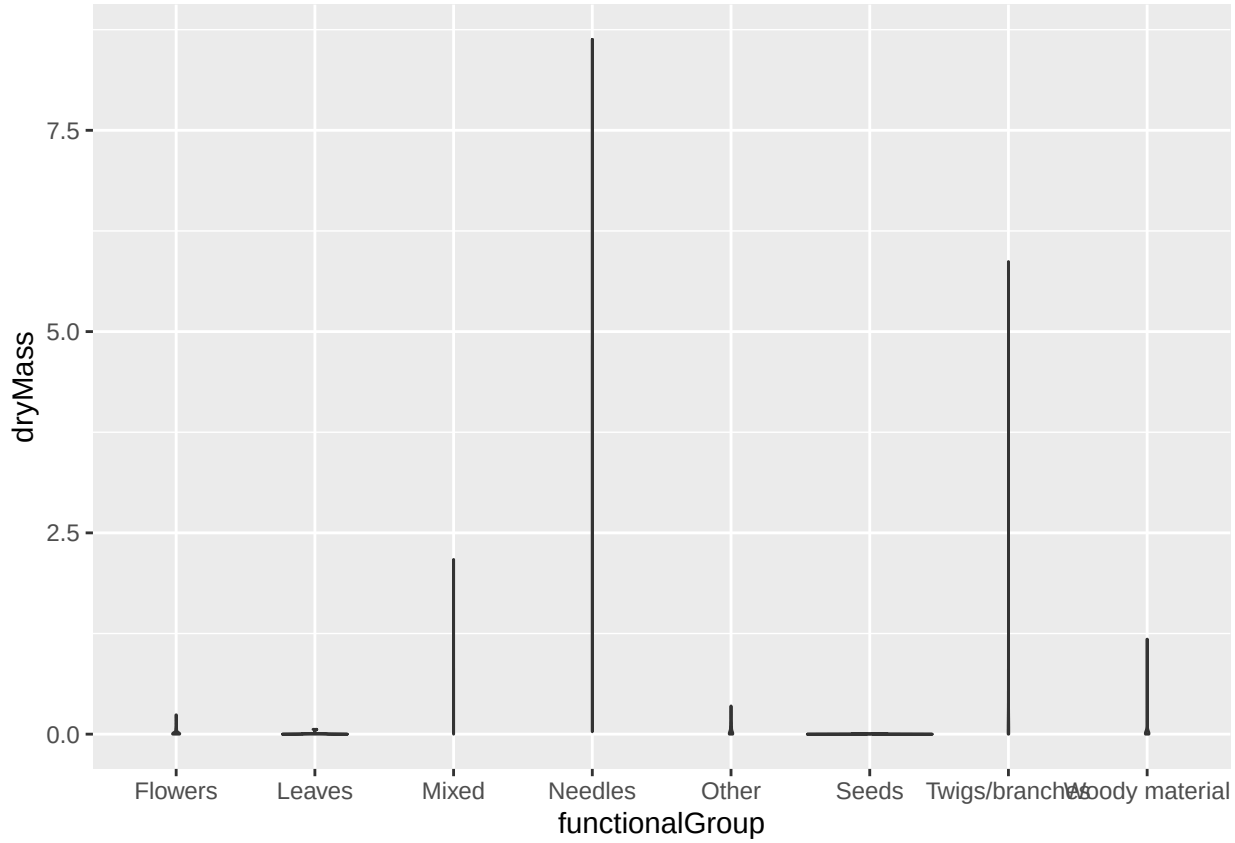


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




```
ggplot(Litter) +
  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 there are not a large amount of multiple measurements at the same dryMass. Because a violin plot is designed to show the range of values and the distribution within that range, the width of the violin is very tiny in this case and results in a poor illistration of the data.

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

> Answer: The boxplot shows that needles and mixed litter tend to have the highest biomass at these sights.