Assignment 3: Data Exploration

Diane Sanchez, 1

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] "/Users/dianesanchez/Documents/Environmental Data Analytics/Environmental_Data_Analytics_2022"
```

```
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: We would be intersted int he ecotoxicology of neonicotinoids becuase we want to know how the pesticides are harming not only the ones that eating and destroying crops, but also on insects such as bees which are vital to pollination and 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: Woody debris & forest litter is an important part of a forest ecosystem becuase it provides multiple ecological services such as carbon sequesetration and helps with soil erosion. But also, too much forest litter can lead to a reduction in sun light going to the soil and can lead to other consequences. Ananlysing the data can help proivde insight in the health of the forest.

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: *Researachers measured data mass based on functinal groups to acount for the differences in weight.* Reserachers used tarps to trap and weigh the litter. The tarps were either randomized or targeted depending on the vegetation. *Reserachers used tower plots. The tower plots are selected randomly within the 90% flux footprint of the primary and secondary airsheds

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim.data.frame(Neonics)
```

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

```
colnames(Neonics)
```

```
##  [1] "CAS.Number"                       "Chemical.Name"
##  [3] "Chemical.Grade"                   "Chemical.Analysis.Method"
##  [5] "Chemical.Purity"                  "Species.Scientific.Name"
##  [7] "Species.Common.Name"              "Species.Group"
##  [9] "Organism.Lifestage"               "Organism.Age"
## [11] "Organism.Age.Units"               "Exposure.Type"
## [13] "Media.Type"                       "Test.Location"
## [15] "Number.of.Doses"                  "Conc.1.Type..Author."
## [17] "Conc.1..Author."                  "Conc.1.Units..Author."
## [19] "Effect"                           "Effect.Measurement"
## [21] "Endpoint"                         "Response.Site"
## [23] "Observed.Duration..Days."         "Observed.Duration.Units..Days."
## [25] "Author"                           "Reference.Number"
## [27] "Title"                            "Source"
## [29] "Publication.Year"                 "Summary.of.Additional.Parameters"
```

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: The effects are specifically of intersts becuase researchers would want to know how the insects react to the different type of pestrices beause based on the harm the pesticides can be baned or if not too harmful can be used.

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, maxsum = 6)
```

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

Answer: They are all bee, they are of most interest becuase they are pollinators and are an important part of the ecosystem.

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

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

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

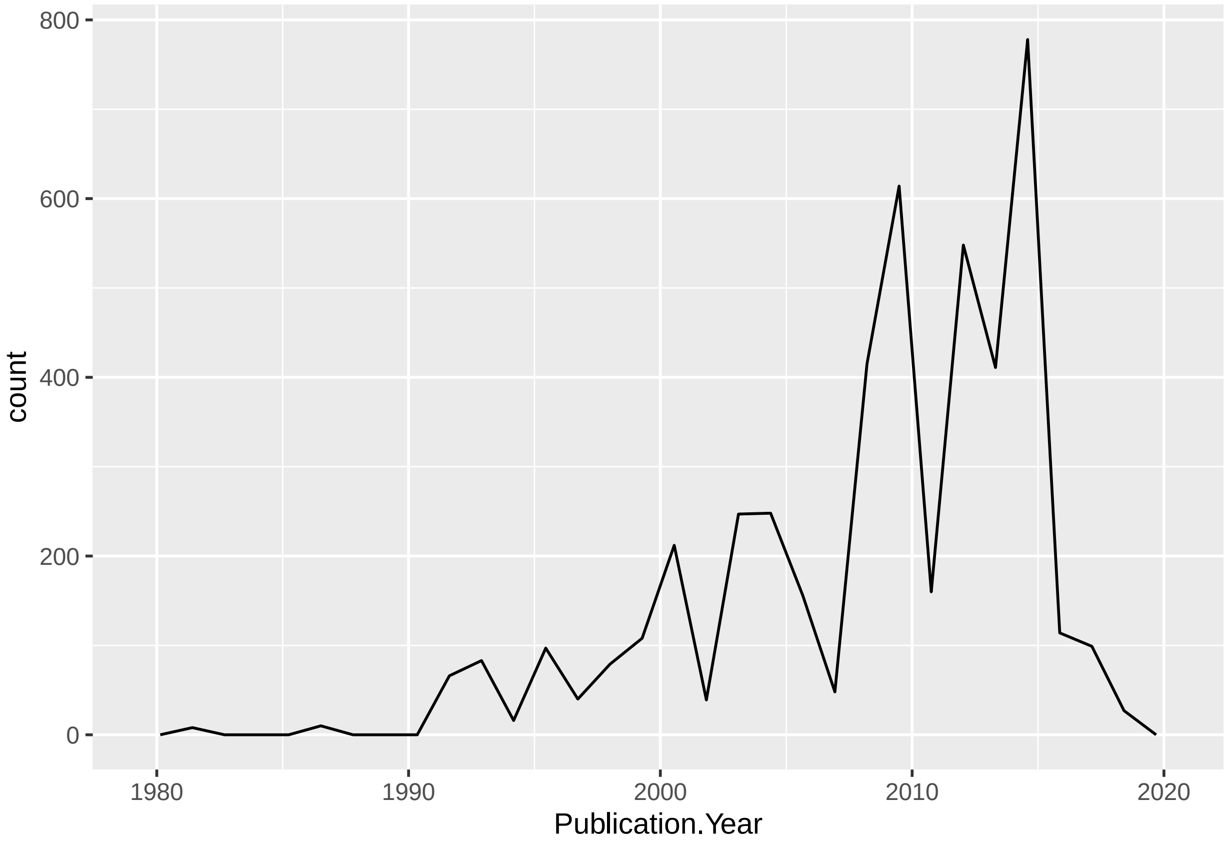
Answer: This is a factor becuase it is a descriptive and it takes less space when storing data. Especially in large data sets, it is easier to navigate and download when you are using the data.

Explore your data graphically (Neonics)

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

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

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

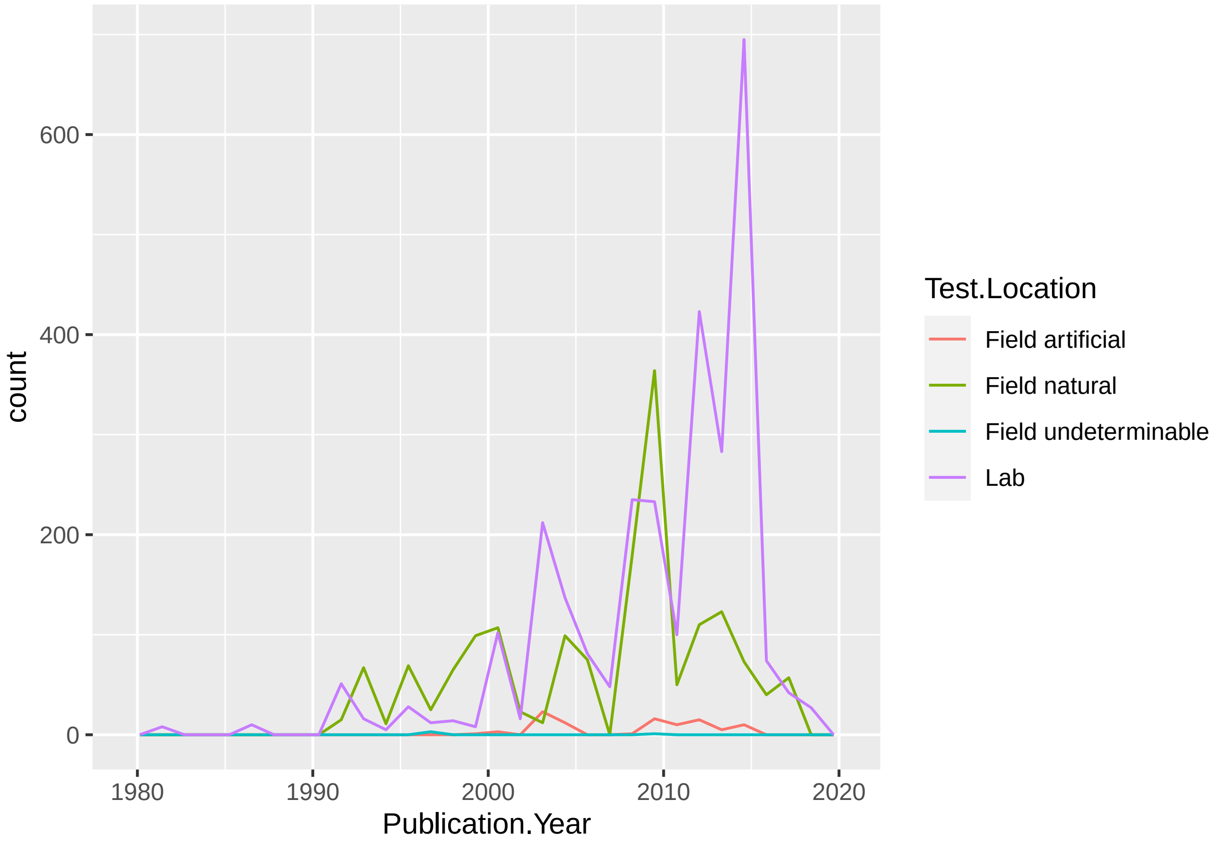


10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
ggplot(Neonics)  +
geom_freqpoly (aes(x = Publication.Year,color = Test.Location, bins = 20 ))
```

```
## Warning: Ignoring unknown aesthetics: bins
```

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

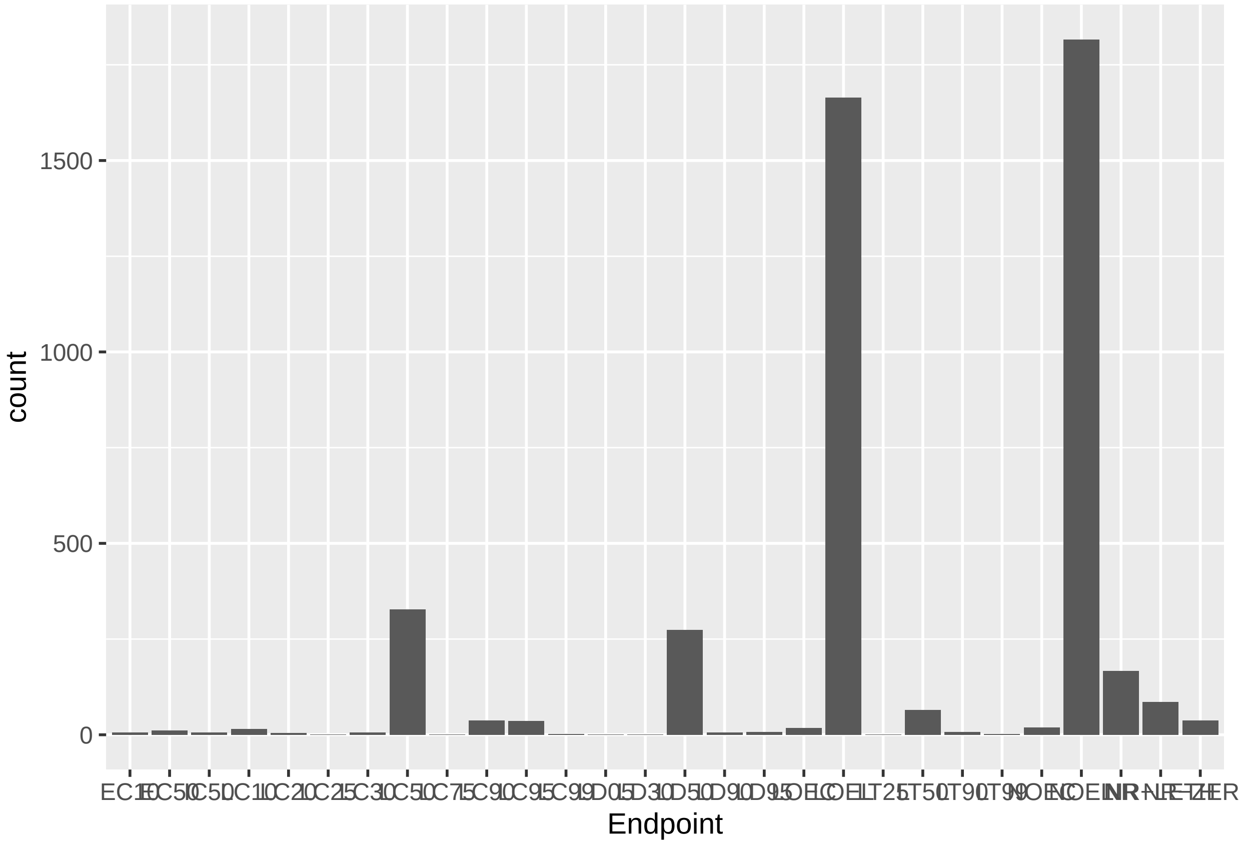


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: the most common test locations are the lab and the natural field. It make sense that most of the testing is completed at the lab and there are lower numbers out in the fields since the testing isn't done in the field.

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, aes(x = Endpoint)) +
  geom_bar()
```



Answer: NOLE and LOLE. NOLE is defined as no-observable-effect-level: highest doese producing effects not significantly different from responses of controls according to author's reported statistical test. LOLE is defined as lowest obersrvable effect level: lowest dose producing effects that were significantly differnect from responses 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)
```

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

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

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

```
unique(Litter$collectDate)
```

```
## [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)
```

```
##  [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
```

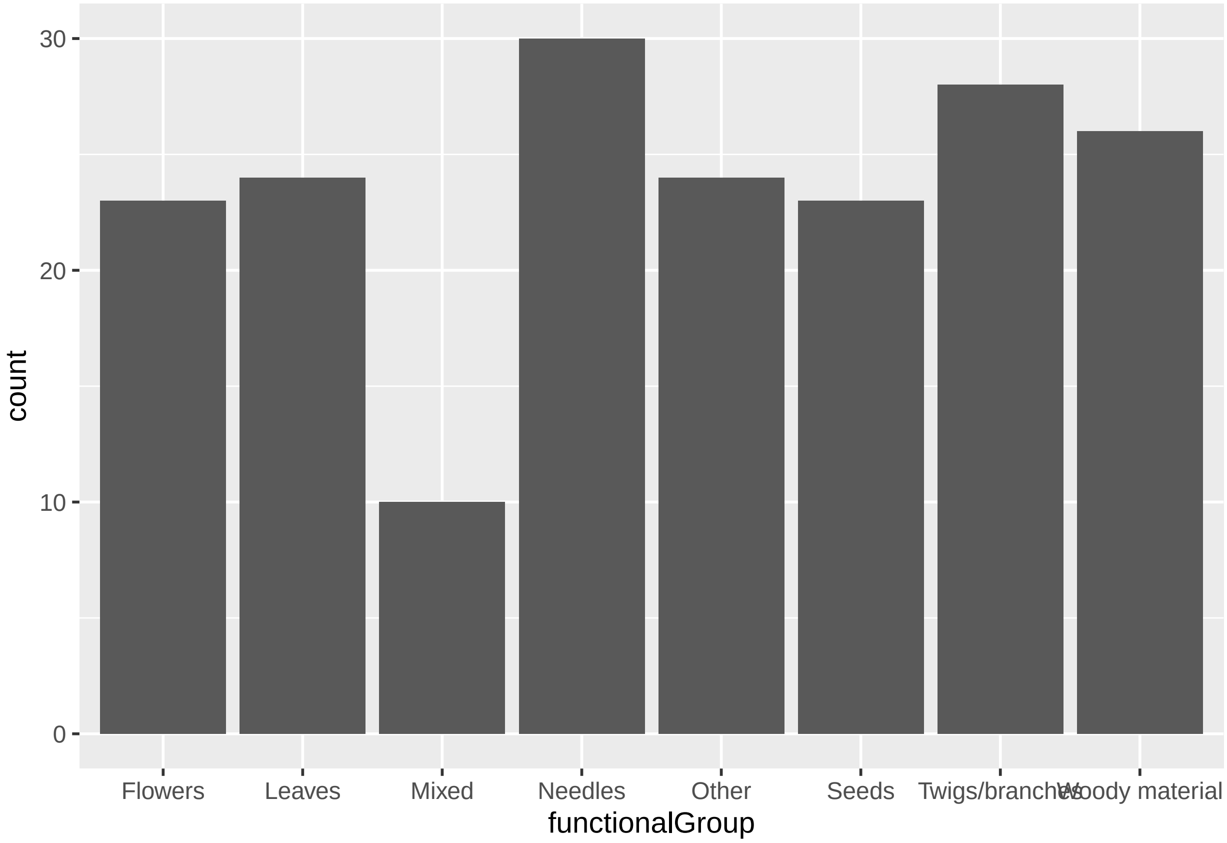
```
summary(Litter$plotID)
```

```
## 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 provided the names of all the plotIDs while on the summary function it provided both the names and how many times each name occured in the data set.

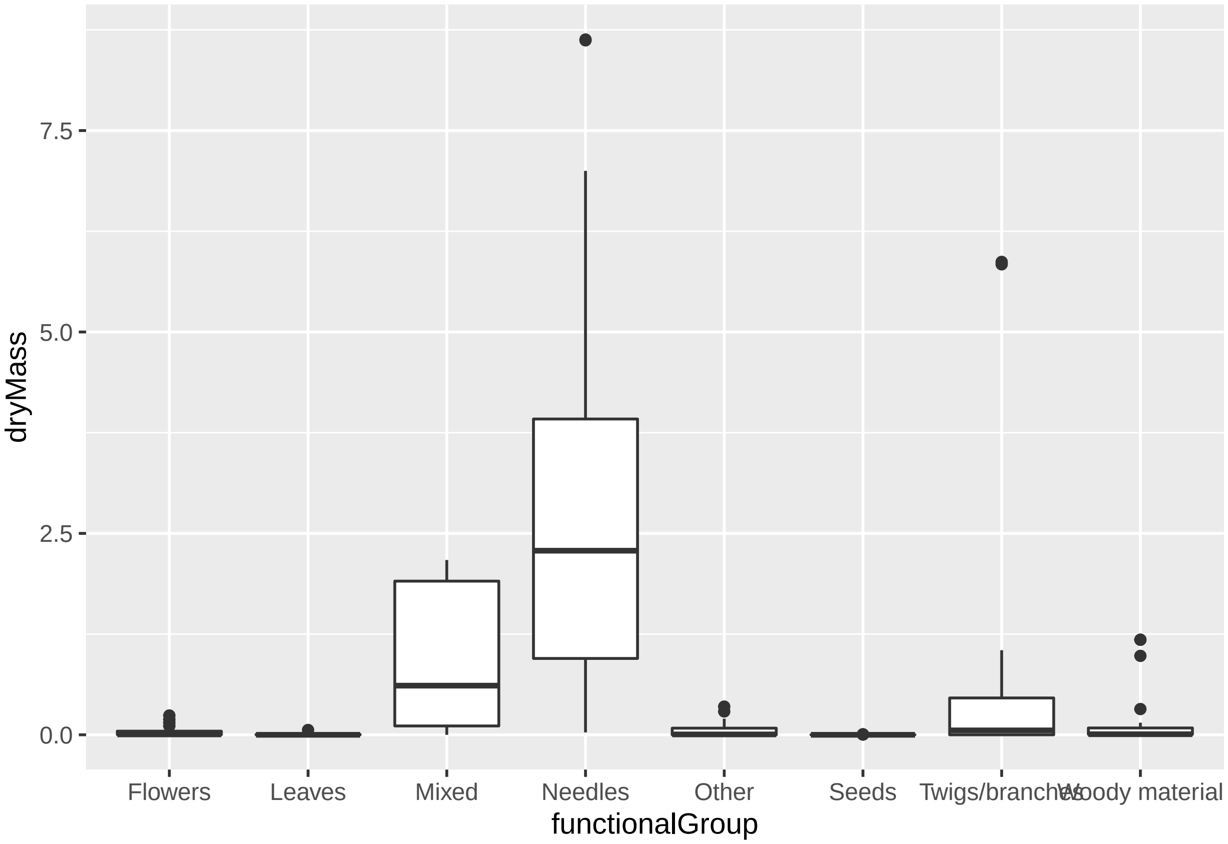
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

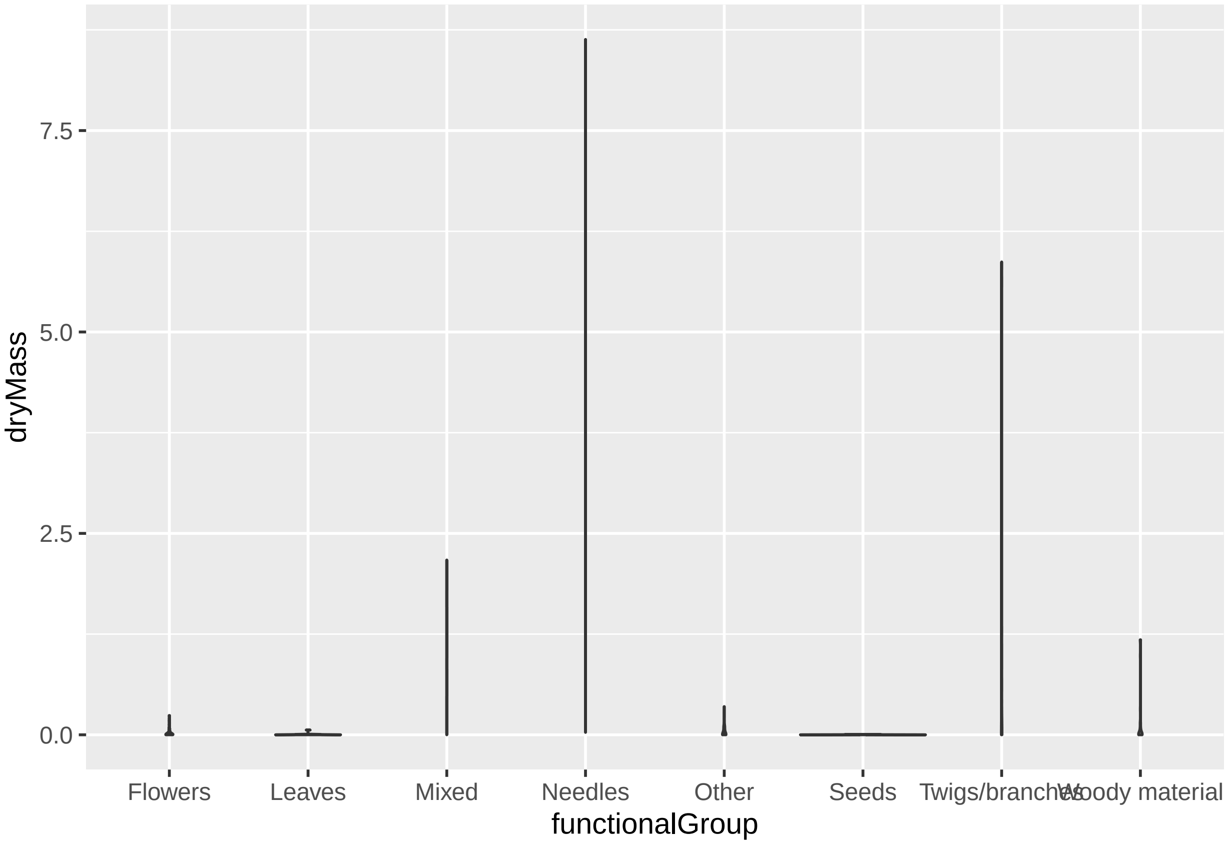


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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



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



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

> Answer: The boxplot is more effective visulization because it gives up a beter visual with more detail on hotspots. There is also not a lot of distribution or heavy contration of mass in the data so the violin looks like a flat line.

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

> Answer: Needles and mixed litter have the highest biomas.