Assignment 3: Data Exploration

Meilin Chan, Section #3

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] "C:/Users/meili/OneDrive - Duke University/EDA/Environmental_Data_Analytics_2022"
```

```
library(tidyverse)


Neonics <- read.csv("../Environmental_Data_Analytics_2022/Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv
                    stringsAsFactors=TRUE)

Litter <- read.csv("../Environmental_Data_Analytics_2022/Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw
                   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: Insects play an important role in their ecosystem and even in agriculture. Pollinators are important to agriculture, and some insects act as an important link in an ecosystem's food web.

If the neonicotinoids have high ecotoxicity on insects, then its use in agriculture would pose a threat to nearby insect species health and the health of the local ecosystem overall.

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 hold a lot of information in terms of forest health (or individual tree health). Researchers can understand the productivity of trees from the nitrogen or carbon content of their leaf litter. Litter data can then be used to see variations in forest productivity, especially with a changing and unpredictable climate.

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:
   The debris is first collected and sorted/identified into categories (ie: leaves, twigs, needles, seeds, flowers, etc.). These sorted groups are then weighed to the accuracy of 0.01 grams (mass data collection). *The litter is collected in ground traps as well as in elevated litter traps (depending on the length and diameter of the material).

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) #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?

```
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 most common effects of the pesticide is on population and mortality. This is because the pesticide product being studied is designed to kill and/or impact the population of whatever is considered a pest to agricultural crops (ie: insects)

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: The most commonly studied species are bees (and the Parisitic Wasp). These species are all well-known pollinators. Understandably, they are of interest over other insects as pollinators are crucial to the productivity of the agricultural sector. If the pesticide being used for agriculture is harming pollinators that support the health and productivity of agricultural crops, then the industry may need to find a safer alternative.

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..Author.)
```

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

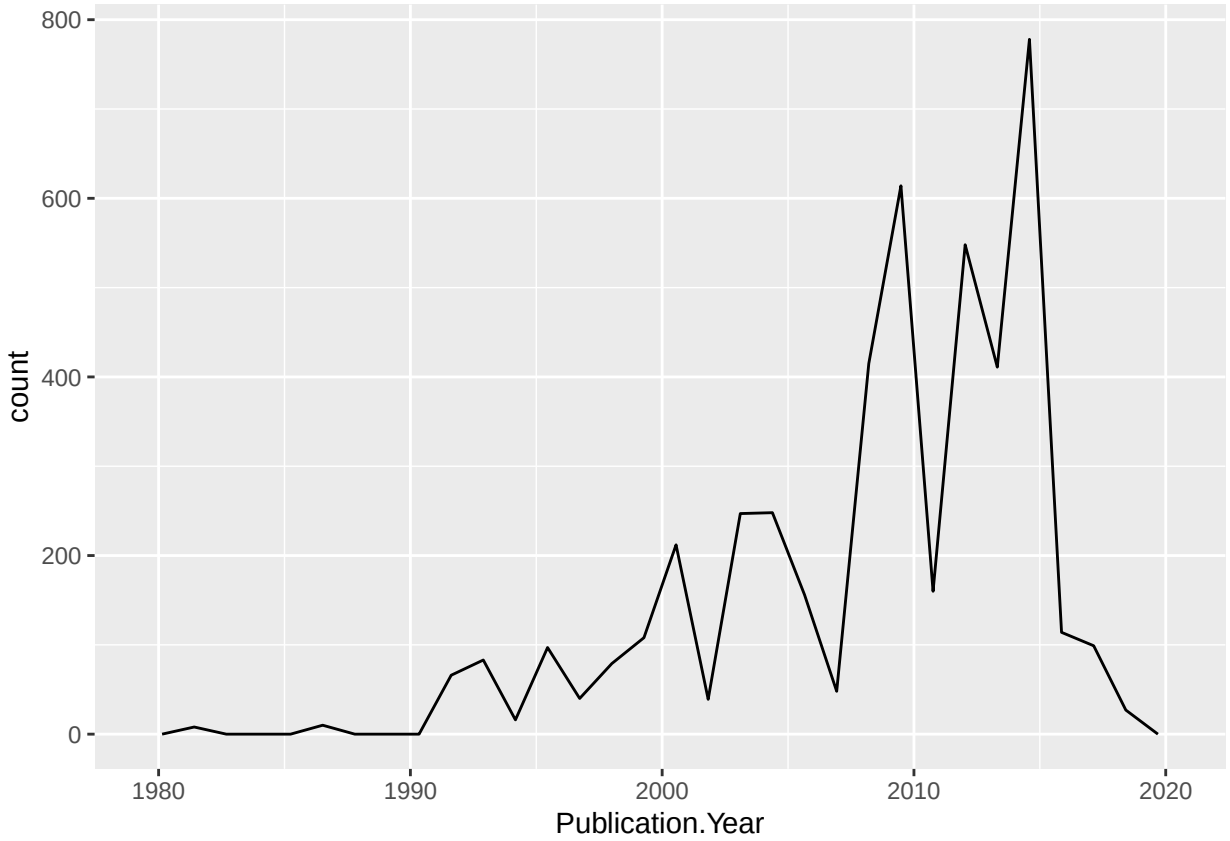
Answer: The class of conc.1..Author is a factor. A factor represents categorical data (many "levels" within a factor. In this dataset, each individual concentration is considered a separate level with the conc.1..author factor. I'm not sure why R would class Conc.1..Author as a factor rather than numeric. My guess is that these concentration values are set to associate with other concentration factors in different columns listed in the data set. So when R processes the data, it sees these links and determines that Conc.1..Author is a factor (categorical).

Explore your data graphically (Neonics)

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

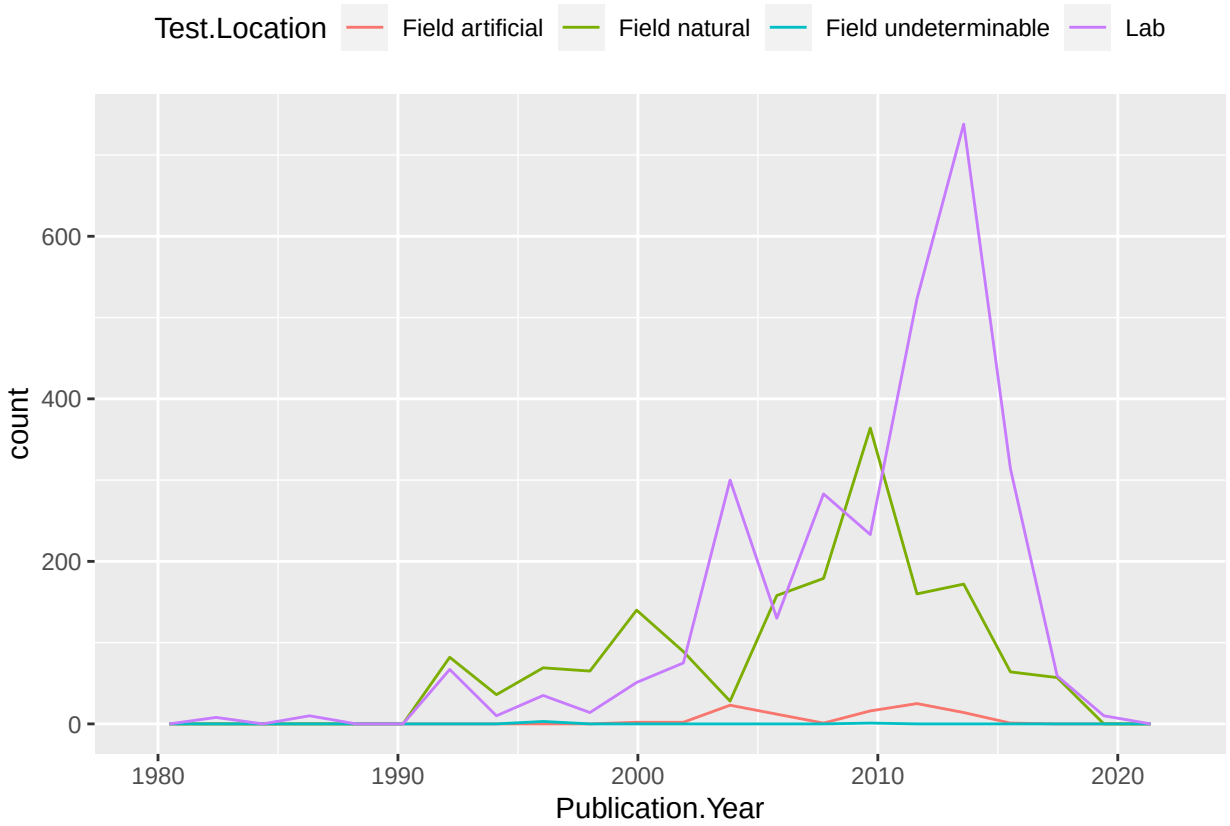
```
ggplot(Neonics) +
  geom_freqpoly(aes(x=Publication.Year))
```

```
## `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) +
  theme(legend.position = "top")
```

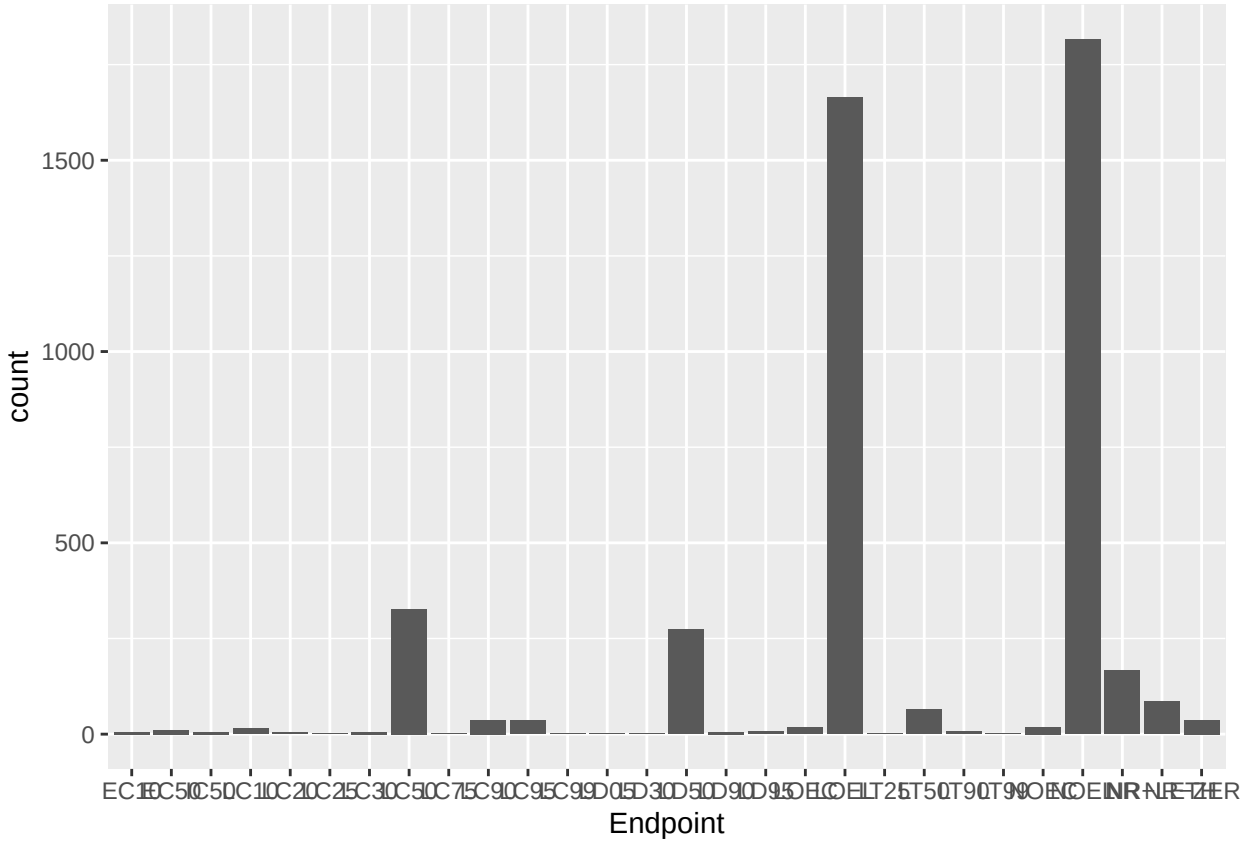


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations look to be in the lab with this test site peaking in the 2010s. The next most common location is in the natural field - the occurence of studies in the natural field is inverse of occurences when studies were conducted in the lab. Lab testing sites were utlizied more from 1990-2000. We can see between 2000-2010, test in the lab began to surpass tests in the natural field with lab testing sites becoming more of the "norm" in the 2010s and beyond.The number of test in general increased significantly from the 2000s to the 2010s.

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 - Lowest-Observable-Effect-Level, or lowest dose producing effects significantly different than author's control, and NOEL - No-Observable-Effect-Level or the highest dose produces effects not significantly different from author's 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) #factor
```

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

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

class(Litter_date)
```

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

```
Aug_2018 <- unique(Litter$collectDate)
Aug_2018
```

```
## [1] 2018-08-02 2018-08-30
## Levels: 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`?

```
Niwot_Plots <- unique(Litter$plotID)
Niwot_Plots
```

```
##  [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
```

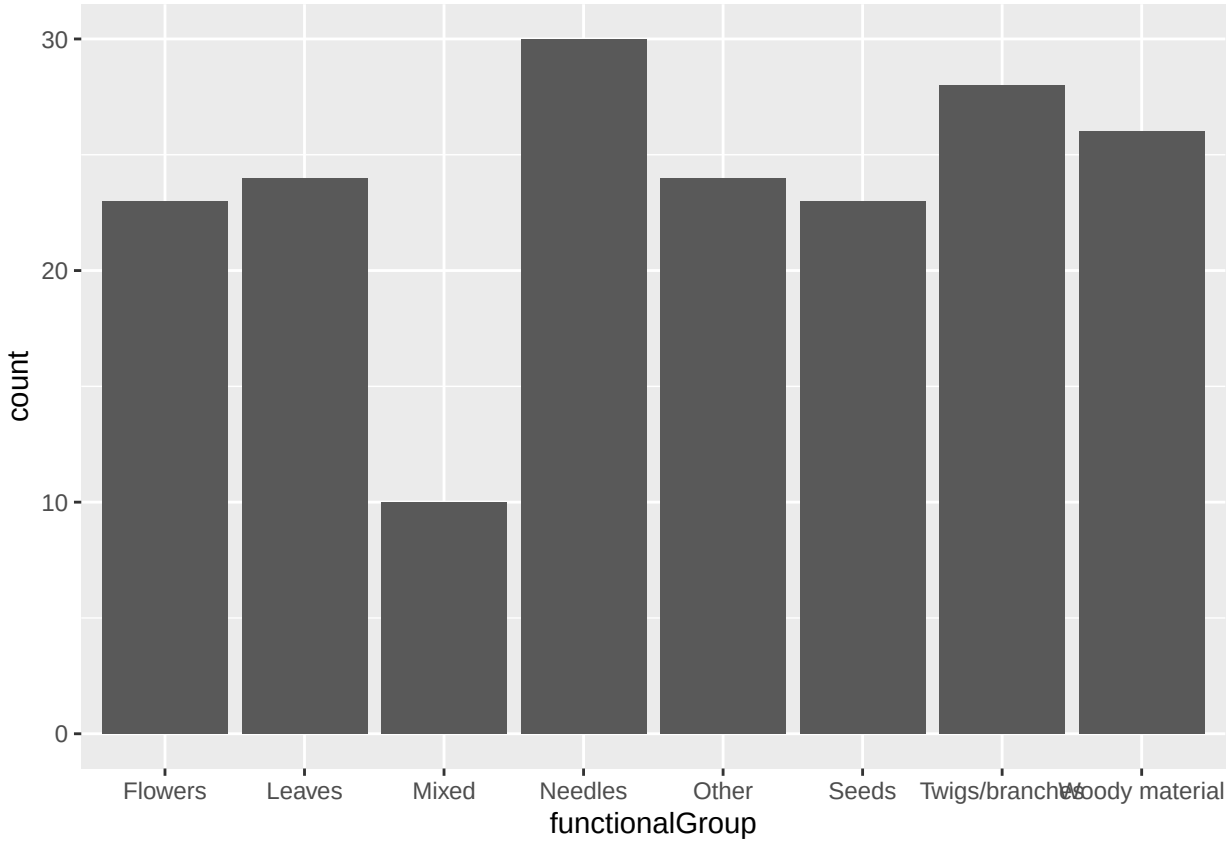
```
summary(Litter$plotID)
```

```
## 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: Summary would tell you how many samples were taken from each plot whereas the unique function would show you the plots that were sampled from without any information on how many samples were taken per plot.

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(functionalGroup))
```

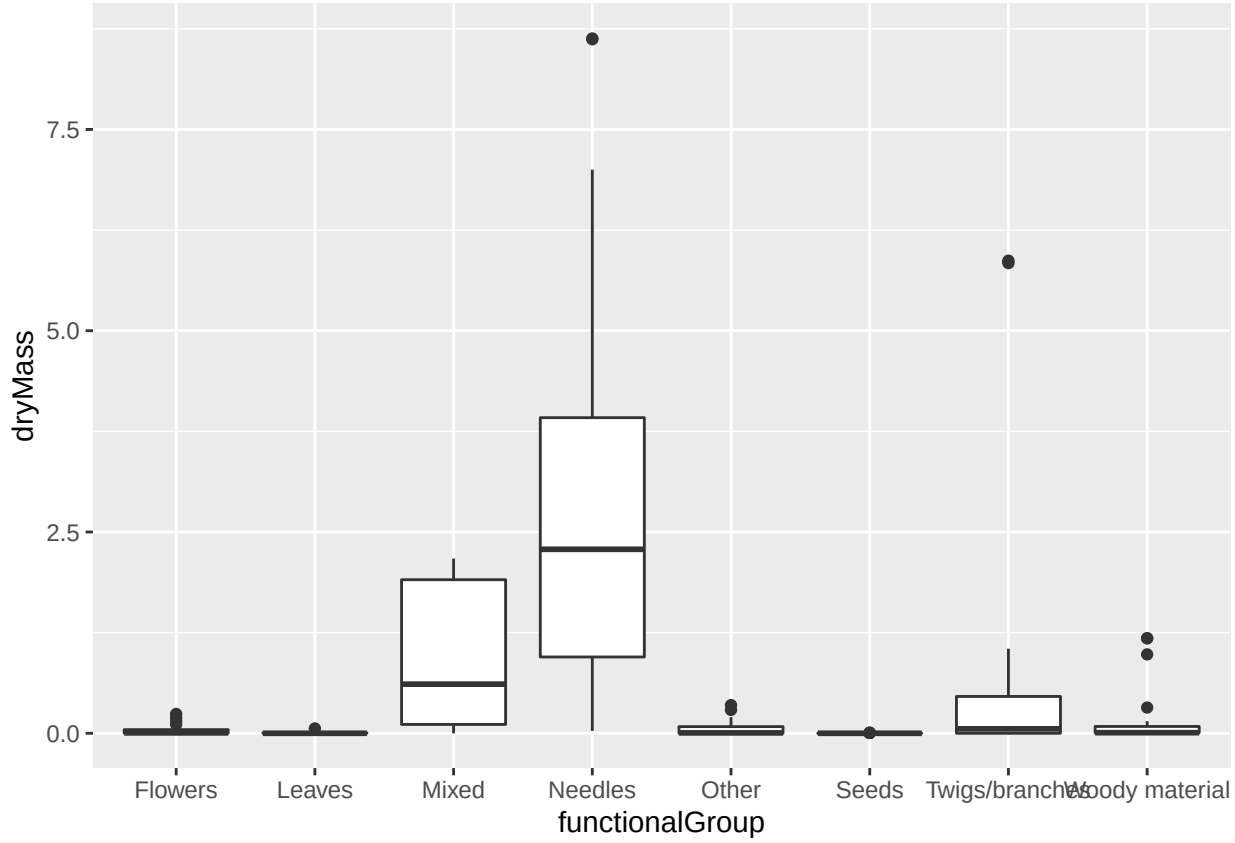


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

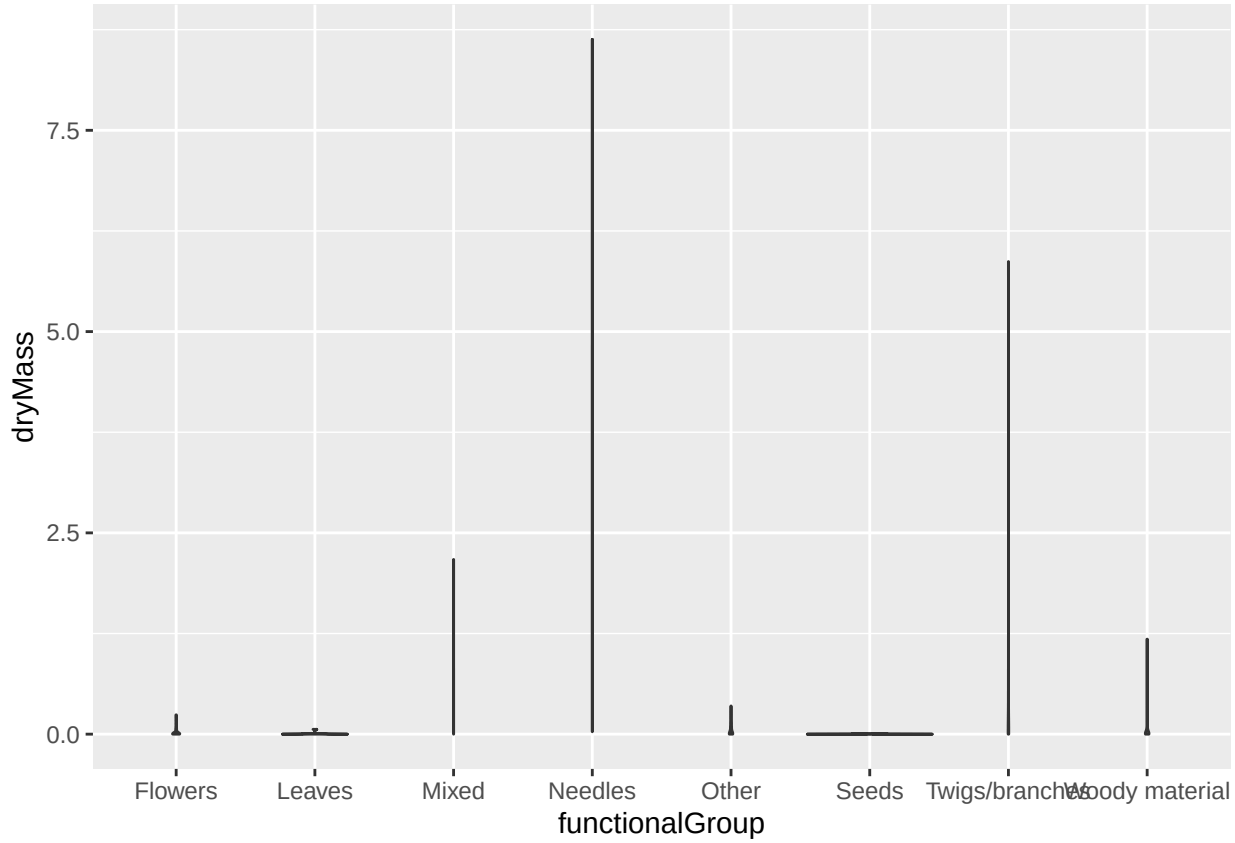
```
Litter_box <- ggplot(Litter) +
  geom_boxplot(aes(x=functionalGroup,y=dryMass))
```

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

Litter_box
```




```
Litter_violin
```



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

> Answer: Each box's width in a violin plot is based on the amount of data at each slice per category portrayed on the x-axis. The violin plot shows the full distribution of data, whereas a box plot only shows min, max, mean, and outliers. However, in this dataset's case, the boxplot isa more effective visualization of the chosen data as the Litter dataset does not have a large enough distribution of data for the violin plot to be effective.

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

> Answer: Needles tend to have the highest biomass at the sample sites.