Assignment 3: Data Exploration

Jiahuan Li

Spring 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.

```
# install.packages("tidyverse")
# install.packages("lubridate")

# check directory
getwd()
```

```
## [1] "d:/Users/Lijh/Desktop/872 R & data analytics/ENV872"
```

```
# upload datasets
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 ecotoxicology of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: That is because if we want to widely apply certain insecticides, we should be cautious about their environmental impacts. Some insecticides will possibly cause severe but unexpected harm which even overweights their benefit. And ecotoxicology is an important perspective to evaluate the potential risks of neonicotinoid insecticides. Actually, there is literature pointing out that neonicotinoid Insecticides could be a major threat to pollinating insects such as bees. They may eliminate not only pests but also pollinators through similar biochemical processes.

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: That is because litterfall in terrestrial ecosystems represents an important pathway for nutrient return to the soil. The data products from litterfall and fine woody debris sampling can provide mass data for plant functional groups. Furthermore, data on litterfall and fine woody debris can be used to calculate annual Aboveground Net Primary Productivity (ANPP) and aboveground biomass at plot, site, and continental levels. They can also provide critical information for understanding the evolution of vegetative carbon fluxes.

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 sampling is planned to take place in 20 40m x 40m plots at sites with forested tower airsheds. And it is designed for 4 40m x 40m tower plots and 26 20m x 20m plots in sites with low-statured vegetation over the tower airsheds. For every 400 mˆ2 plot area, one litter trap pair (one elevated trap and one ground trap) is deployed, resulting in 1-4 trap pairs per plot.
   2. Depending on the vegetation, trap placement within plots can be either targeted or randomized. Litter trap placement is random in sites with >50% aerial cover of woody vegetation >2m in height.
   3. Ground traps are sampled once per year. The frequency of target sampling for elevated traps varies depending on the vegetation present at the site, with frequent sampling (1x every 2 weeks) in deciduous forest sites during senescence and infrequent year-round sampling (1x every 1-2 months) at evergreen sites.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dimensions = dim(Neonics)

# Dataset dimensions: rows = 4623 and columns = 30
print(paste("The dimension of data frame is:", dimensions[1], dimensions[2]))
```

```
## [1] "The dimension of data frame is: 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?

```
sort(summary(Neonics$Effect), decreasing = TRUE)
```

> Answer: Population and mortality are the most common effects with the frequency of 1803 and 1493, respectively. That may be because they are closely related to the potential environmental harm of insecticides. Scientists may focus more on the changes of population abundance and mortality rate of certain insect groups, which could be straightforward and important indicators for the environmental impacts caused by the insecticides.

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...]

```
sort(summary(Neonics$Species.Common.Name), decreasing = TRUE)
```

> Answer: The six most commonly studied species are Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. These species have high genetic similarity and all belong to Apidae according to biotaxonomy. They are of special interest because they are regarded as the best pollinators of crops around the world. Thus, the declining population of these important pollinators day by day is a major threat to the agriculture. And according to the literature, the use of neonicotinoid pesticide is considered to be the major reason for the decline of bees in the world.

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?

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

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

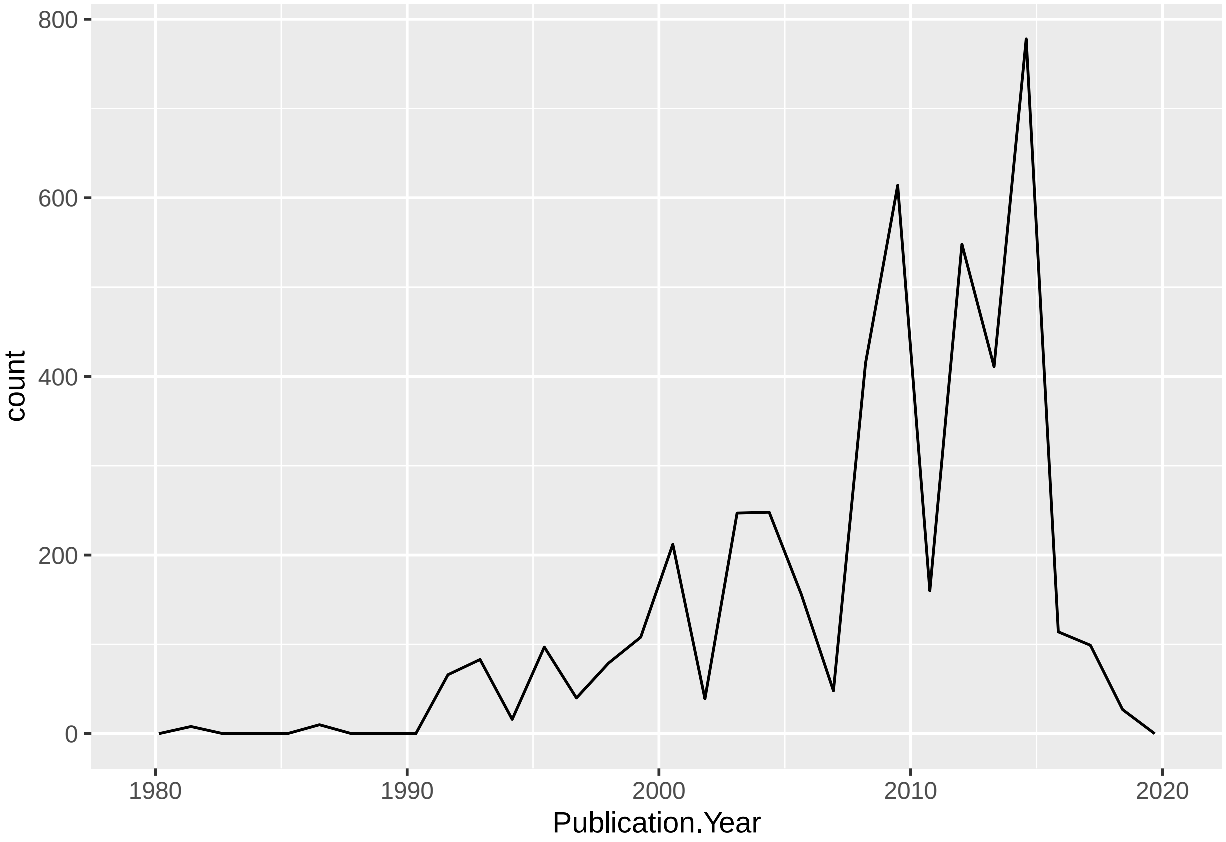
> Answer: the class of this column is factor. It is not numeric because this column is not purely composed of numeric values. Some values in the column are not numbers and they are viewed as strings.

Explore your data graphically (Neonics)

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

```
library(ggplot2)
ggplot(Neonics, aes(Publication.Year)) + geom_freqpoly()
```

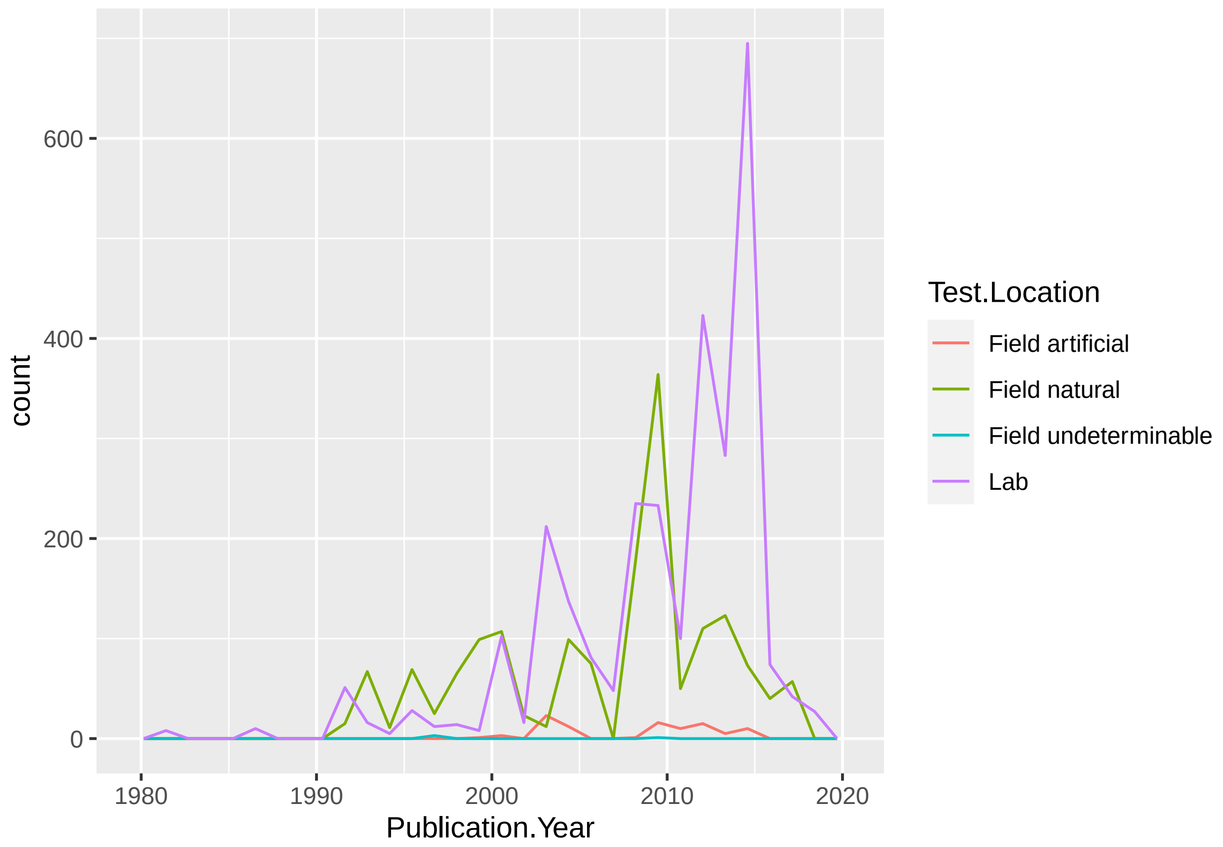
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

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

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



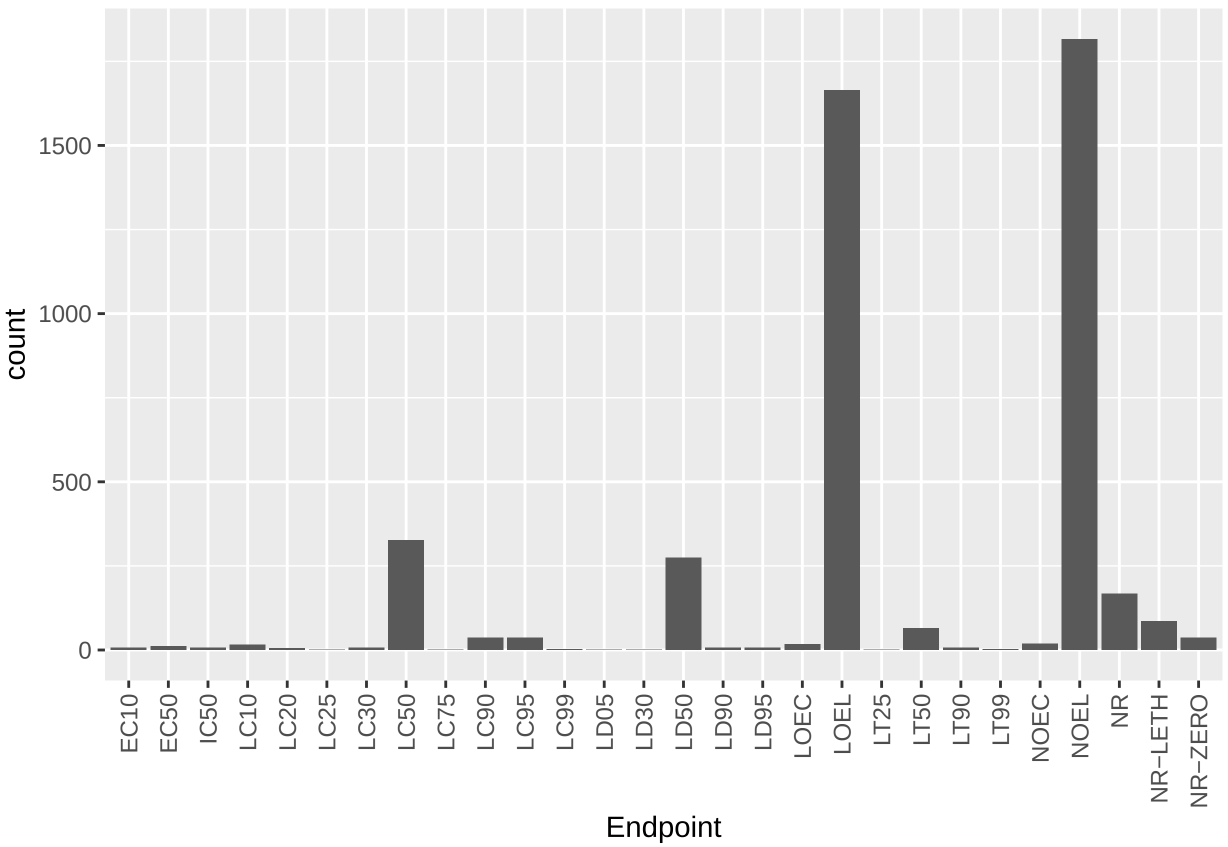
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test location is lab for most of the time. And papers with natural field as test locations exceeded the number of lab ones over the periods of 1992 - 2001 and 2008-2011.

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(Neonics, aes(Endpoint)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: the two most common end points are LOEL and NOEL. The definition of LOEL (Lowest-observable-effect-level) is the lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC). And the definition of NOEL (no-observable-effect-level) is the highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test (NOEAL/NOEC).

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"
```

```
# the original class is factor, not a date

# convert to the date type
Litter$collectDate <- as.Date(Litter$collectDate)
class(Litter$collectDate)
```

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

```
ans = unique(Litter$collectDate)
print(paste("The unique sampling dates in August 2018 are", ans[1],",", ans[2]))
```

```
## [1] "The unique sampling dates in August 2018 are 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
```

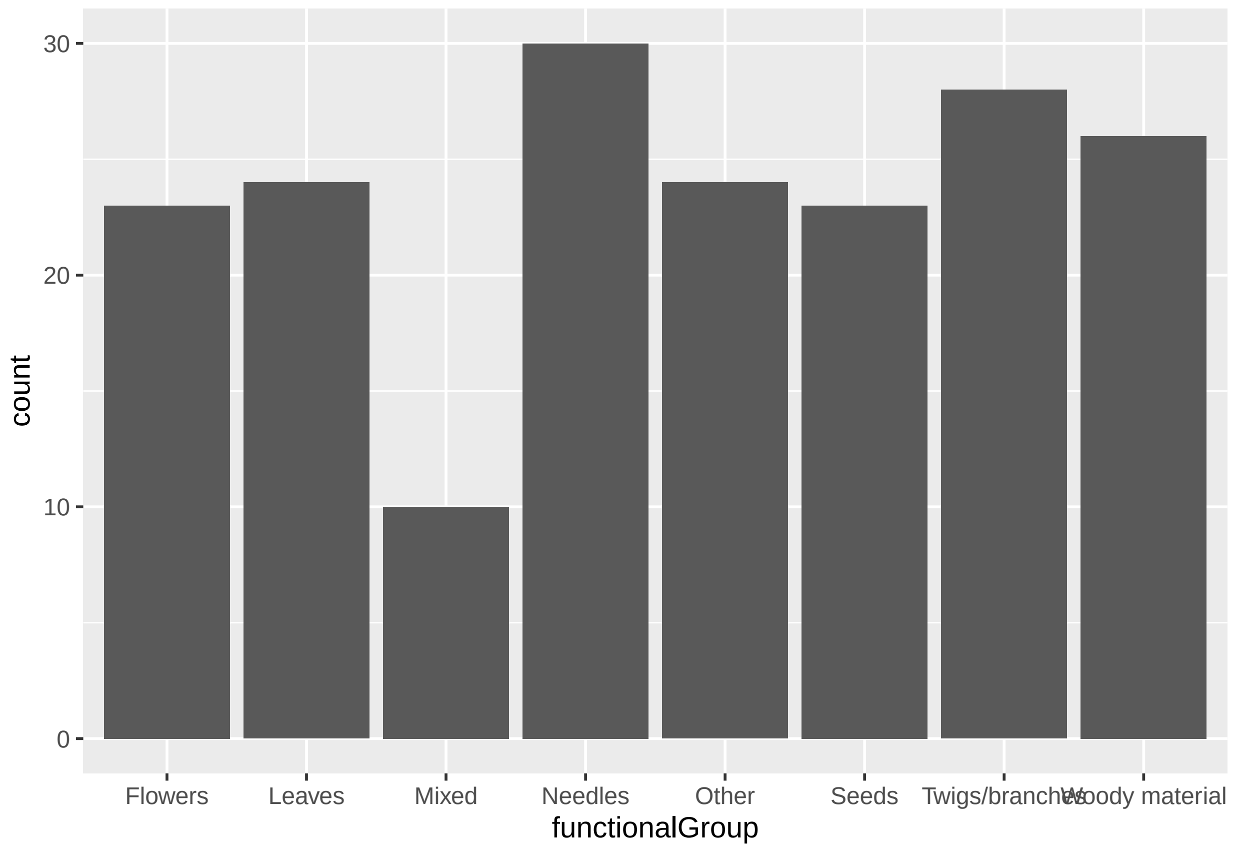
```
summary(Litter$plotID)
```

```
## 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: 12 plots. The `unique` function returns a vector, data frame or array like `x` but with duplicate elements/rows removed. While the `summary` is a generic function used to produce result summaries of the results of various model fitting functions. The information from both `unique` and `summary` contains the list of unique plot names. But `unique` also gives the sum of the list elements. While `summary` returns the sum of all the elements in the dataframe sampled at each of the 12 plots.

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.

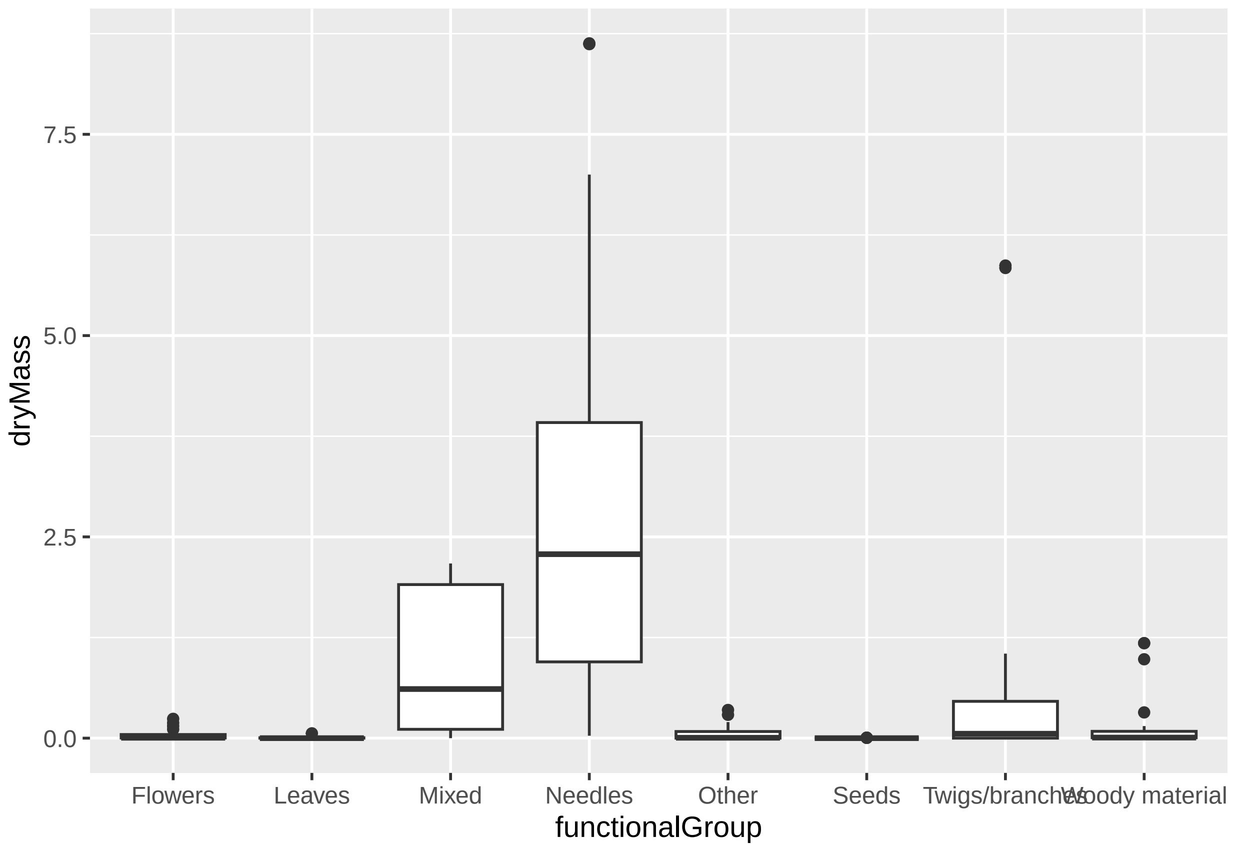
```
ggplot(Litter, aes(functionalGroup)) + geom_bar()
```



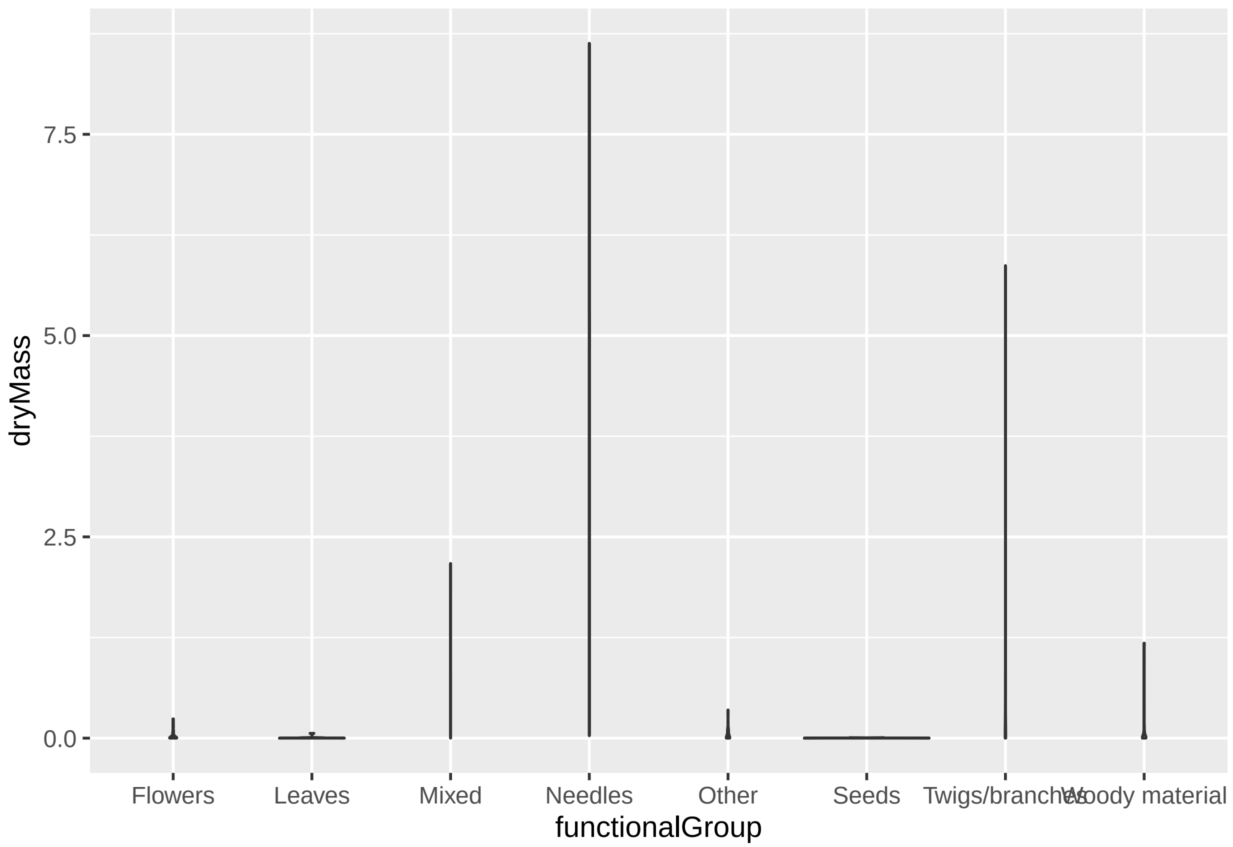
```
# shown in the graph, the litter types are nearly equally distributed.
```

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

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



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



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

> Answer: violin plots are similar to box plots, except that they also show the kernel probability density of the data at different values. However, because the data in this dataframe is sparse, i.e., the observations of function groups on different dryMass are limited and without distinct difference, the width of the violin plot is small and it looks just like a line. Therefore, the shape of the violin plot can hardly be identified. However, for the boxplot, the box width is fixed and its shape is illustrated according to the quantiles of the dataframe. Thus, the boxplot will always keep a recognizable shape.

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

> Answer: Needles