Assignment 3: Data Exploration

Yuxiang Ren

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.

```
#packages
library(tidyverse)
library(lubridate)
library(ggplot2)
library(dplyr)

#datasets:
##Neonics
Neonics <-read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",stringsAsFactors = T)
##Litter
Litter <- read.csv("./Data/Raw/NIWO_Litter/NEON_NIWO_Litter_massdata_2018-08_raw.csv",
                   stringsAsFactors = T)
```

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: Neonicotinoids are one of the most widely used insecticides in agriculture. However, with more and more use, various negative effects of neonicotinoids are gradually being known and researched. First, the impact of neonicotinoids on insect populations could might directly influence the entire ecosystem. Second, some reports show that using these insecticides also declines the population of natural pollinators, especially bees, which could have serious implications for food security. Third, neonicotinoids have been found to be persistent in the environment and could damage surrounding wildlife for a long time. Therefore, research in this field can make people use pesticides more rationally and effectively. This will also benefit the environment.

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: The study of litter and woody debris in forests provides valuable information about the dynamics of the forest floor, the biodiversity of the ecosystem, and the impact of forest management practices.

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 and fine woody debris are sampled at terrestrial NEON sites where the woody vegetation is taller than 2 meters. 2. At sites with forested tower airsheds, the litter sampling is focused on 20 40m x 40m plots 3. Ground traps are sampled once per year.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

```
# Neonics has 4623rows, and 30ncolumns.
```

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
effect_Neo <- summary(Neonics$Effect)
#sort, and used names to extract the first three effects
mostcommon_effect <- names(sort(effect_Neo,decreasing = TRUE)[c(1)]); mostcommon_effect
```

```
## [1] "Population"
```

> Answer: 4. The most common effect is population. The reason why they are interested in that they provide important information about the impact of these 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...]

```
species_Neo <- summary(Neonics$Species.Common.Name)
sixcommon_species <- names(sort(species_Neo, decreasing = TRUE)[c(1:6)])
print(sixcommon_species)
```

```
## [1] "(Other)"               "Honey Bee"             "Parasitic Wasp"
## [4] "Buff Tailed Bumblebee" "Carniolan Honey Bee"   "Bumble Bee"
```

> Answer: They are "other", Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee and Bumble Bee. What they have in common is that they are all bees, which are important pollinators

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?

```
#data type
str(Neonics$Conc.1..Author.)
```

```
##  Factor w/ 1006 levels "<0.0004","<0.025",..: 639 510 813 622 442 637 500 642 814 784 ...
```

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

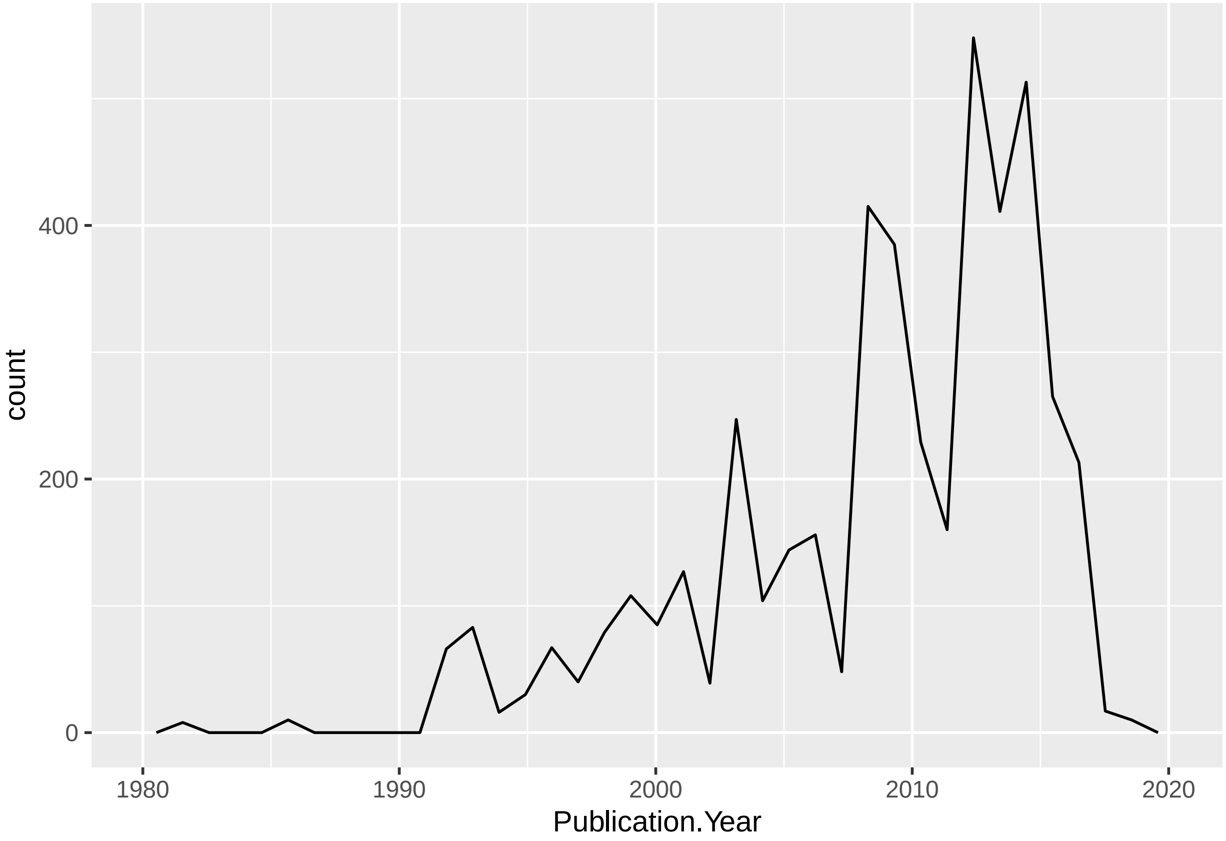
```
##    0.37/      10/      NR/       NR        1     1023    0.40/       2/
##      208      127      108       94       82       80       69       63
##       10   0.053/      100      50/     0.5/     0.03    0.05/     0.45
##       62       59       56       51       45       44       43       43
##     0.1/    0.45/     1.0/    2.27/       50    0.125     500/      0.5
##       42       40       40       40       36       33       33       32
##   0.048/    0.15/       1/       48    25.0/      12/    0.027      2.4
##       30       30       30       30       28       27       26       26
##     0.2/    0.56/     100/        3    0.01/    1000/       3/    0.336
##       25       24       23       23       22       22       22       21
##     1.5/     0.05      1.5    2.60/    20.0/        6    6.80/    62.5/
##       21       20       20       20       20       20       20       20
##    0.005     0.4/    0.18/     0.3/     1000       40 0.00355/      0.1
##       18       18       17       17       17       17       16       16
##      0.4     150/      300      80/    0.053     0.24     0.28     125/ 
##       16       16       16       16       15       15       15       15 
##        9   0.0001  0.0004/   0.084/     0.15      0.6    12.5/   144.0/ 
##       15       14       14       14       14       14       14       14 
##     350/    40.0/      48/       56      84/    0.17/      125       14 
##       14       14       14       14       14       13       13       13 
##       16       17   0.047/    0.25/    0.28/    1.28/    1.81/      112 
##       13       13       12       12       12       12       12       12 
##      150     2.5/       25      60/      75/    0.02/   0.025/     0.29 
##       12       12       12       12       12       11       11       11 
##    37.5/       4/        5  (Other) 
##       11       11       11     1817
```

Answer: The "Conc.1..Author." column in the dataset is a factor variable with 1006 levels. The reason why it is not a numeric parameter is because there are symbols in the column, such as the greater than sign and the less than sign.

Explore your data graphically (Neonics)

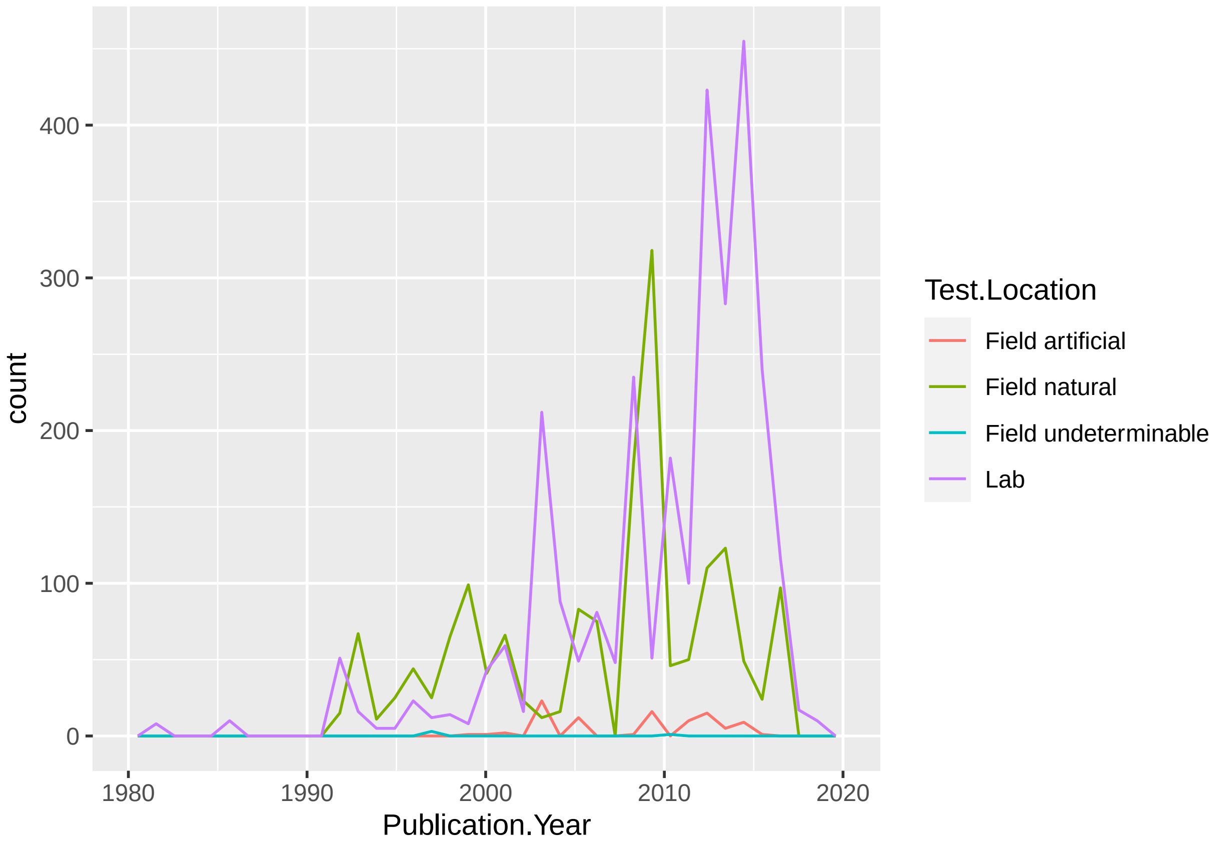
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 37)
```



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

```
ggplot(Neonics, aes(x = Publication.Year, color = Test.Location)) +
  geom_freqpoly(bins = 37)
```



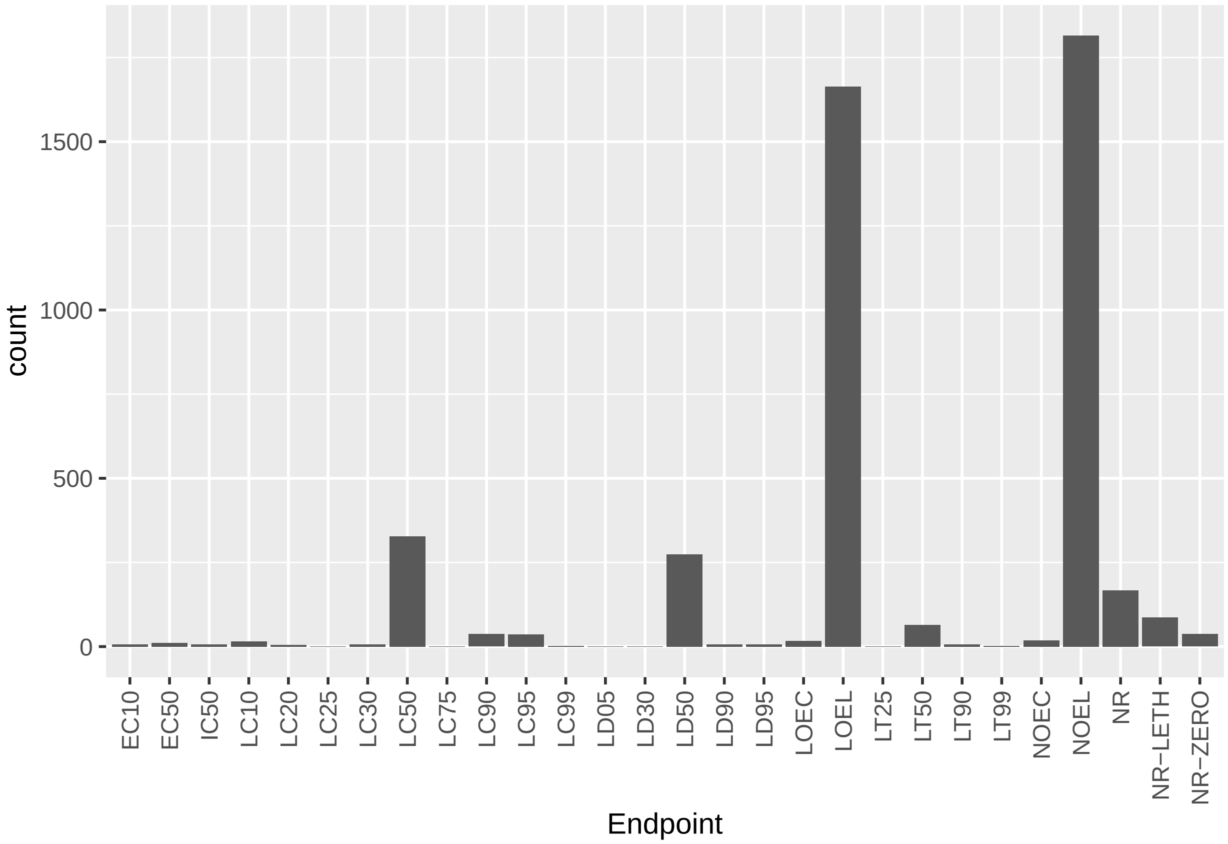
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: Labs are the most common testing location, especially after 2010. From 1990 to 2000, more research like to conduct experiments in the natural field. Between 2000 and 2015, there were a small number of artificial field test sites.

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



```
summary(Neonics$Endpoint)
```

```
##    EC10    EC50    IC50    LC10    LC20    LC25    LC30    LC50    LC75    LC90
##       6      11       6      15       5       1       6     327       1      37
##    LC95    LC99    LD05    LD30    LD50    LD90    LD95    LOEC    LOEL    LT25
##      36       2       1       1     274       6       7      17    1664       1
##    LT50    LT90    LT99    NOEC    NOEL      NR NR-LETH NR-ZERO
##      65       7       2      19    1816     167      86      37
```

Answer: LOEL and NOEL are the two most common end points. LOEL is the lowest concentration at which a toxic effect can be observed and statistically distinguished from normal or control responses. NOEL is the highest concentration at which no adverse effects can be observed, and the effects are not statistically different from normal or control responses.

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$collectDate)
# change to data
Litter$collectDate <- ymd(Litter$collectDate)
# find dates
august_2018 <- unique(Litter$collectDate[format(Litter$collectDate, "%m") == "08"
                                          & format(Litter$collectDate, "%Y") == "2018"])
print(august_2018)
```

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`?

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

```
##  [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
```

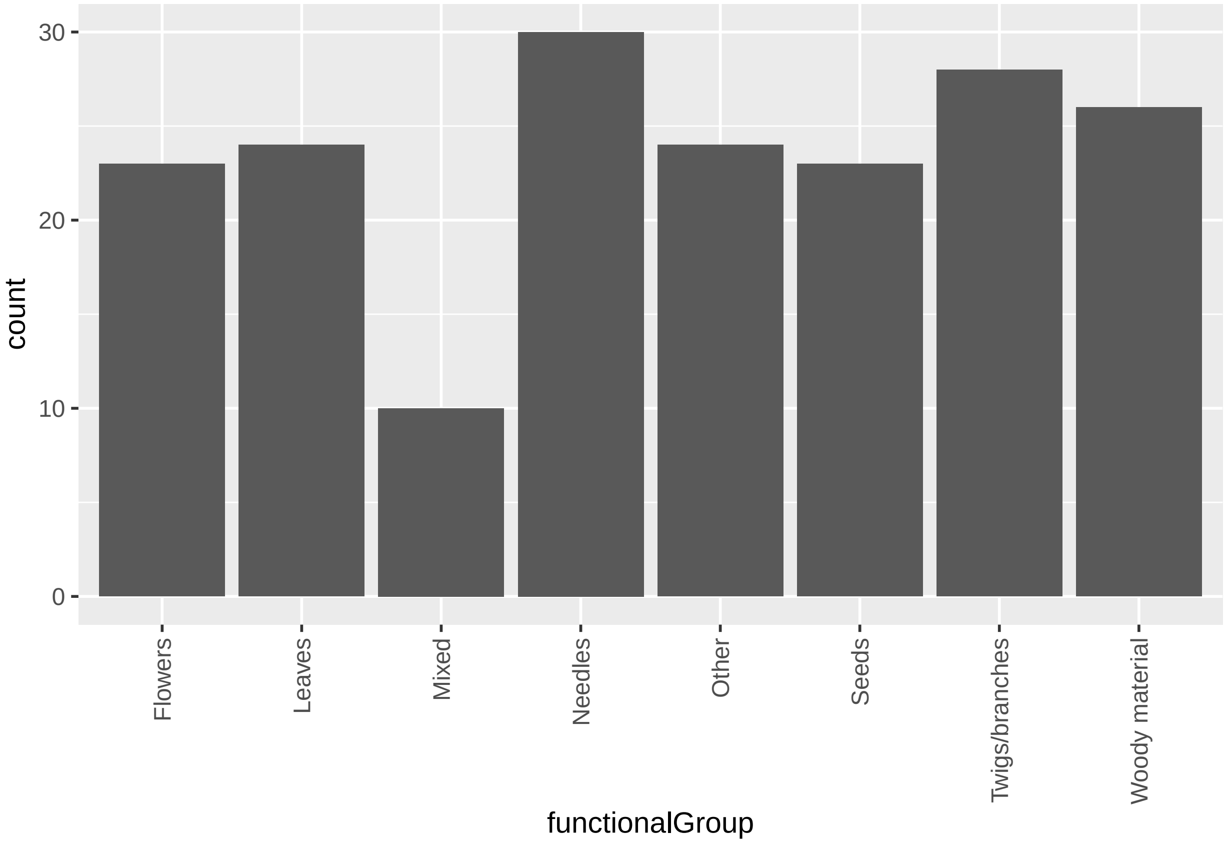
```
summary(Litter$plotID)
```

```
## 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: There are 12 plots sampled at Niwot Ridge. For 'unique', it lists the total number of factors and each factor's name. The summary function would count the number of each factor in the column.

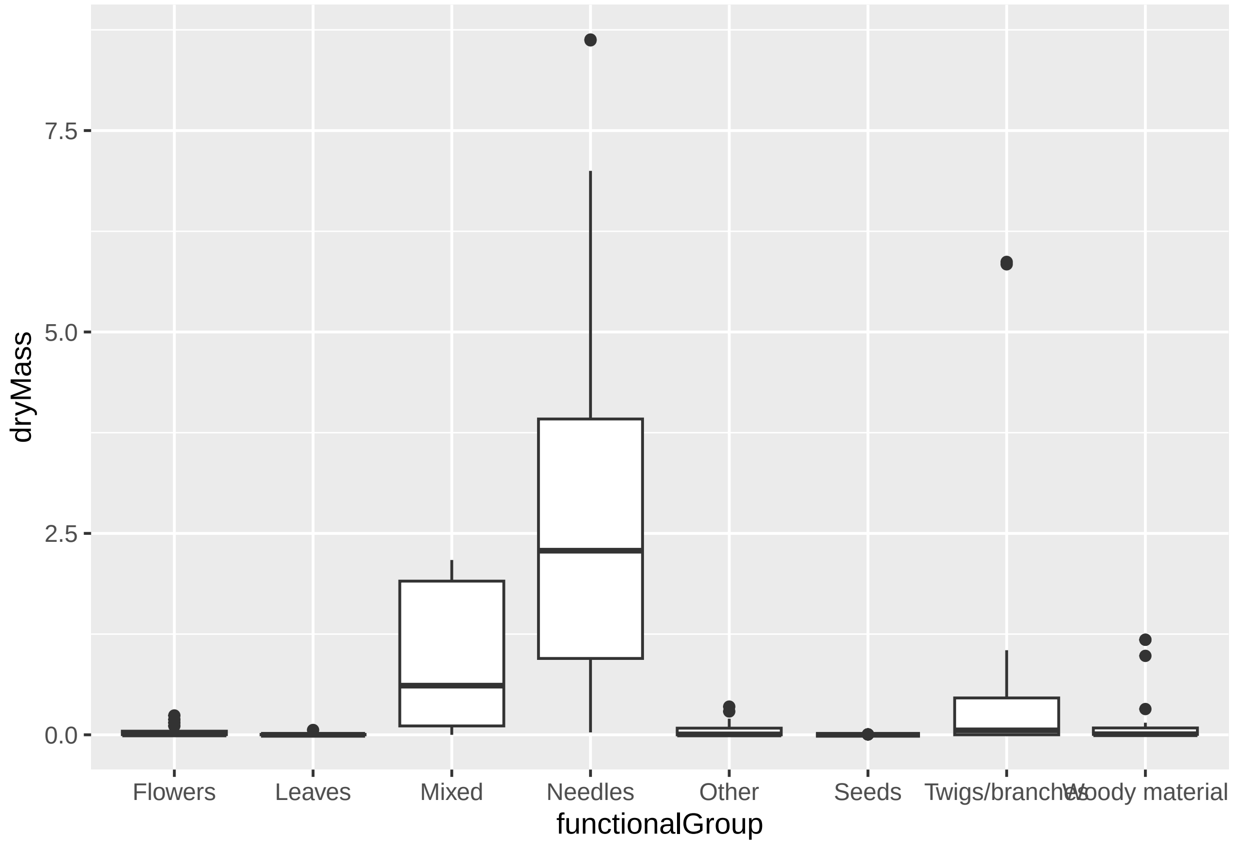
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#bar graph
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar()+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

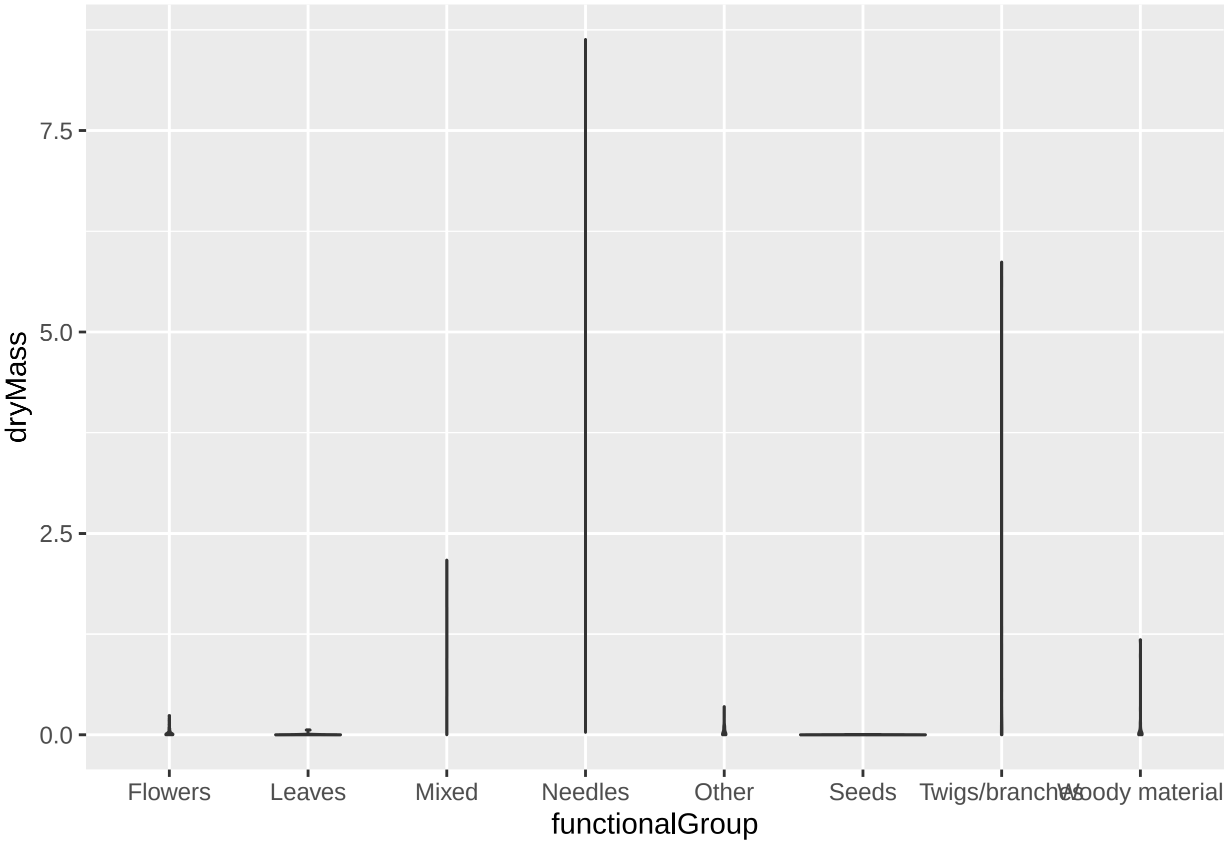


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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



```
#geom_violin
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75))
```



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

> Answer: The violin plot only shows as lines. It cannot show the distribution information of the data. It might be caused by the data being plotted is not dense enough to create a recognizable violin shape. Therefore, a boxplot is a more effective visualization option.

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

> Answer: Needles