# Assignment 3: Data Exploration

Kelly Davidson

## 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. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.
6. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai.

The completed exercise is due on Sept 30th.

### 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 include the subcommand to read strings in as factors.

```
#1. wrap code when knitting
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=80), tidy=TRUE)

# Checking current working directory and load tidyverse package
getwd()
```

```
## [1] "/home/guest/EDA-Fall2022/EDA-Fall2022"
```

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

# Uploading & naming neonics dataset
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = TRUE)

# Uploading & naming litter dataset
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: Since neonicotinoids are a type insecticide, we would likely be interested in studying the effects of these insecticides on many different types of insects. Certain insects such as bees are important pollinators and are a necessary part of food production and agriculture, so we would want to know the effect of these neonicotinoids on all insects to determine overall impacts on target and non-target species.

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: Since leaf litter and woody debris are key aspects of forest ecosystems, we would likely be interested in analyzing this information to study forest health. Leaf litter and woody debris both play important roles in nutrient cycling. As they decompose, they release nutrients into the soil and also help retain moisture.

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. Sampling occurs at terrestrial NEON sites that contain woody vegetation greater than 2 meters tall 2. Sampling occurs only in tower plots - the location of which are selected randomly 3. The size of sampling plots vary depending on the type of vegetation present (forested vs. low-statured vegetation)

## Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
# 5. Determining the dimensions of the Neonics dataset

dim(Neonics)
```

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

```
# 4623 rows and 30 columns
```

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?

```
# 6. Using the 'summary' function on the 'Effect' column to review the most
# common effects that are studied

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 that are studied are population (count = 1803) and mortality (count = 1493). These effects might be of interest specifically to analyze the impact of certain neonicotinoids on target and non-target species. Additionally, understanding if a specific neonicotinoids results in death (mortality) or not would be essential in determining the effectiveness of the insecticide.

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.

```
# 7. Using the 'summary' function to determine the 6 most commonly studied
# species (common name) in the dataset

summary(Neonics$Species.Common.Name)
```

```
##                          Honey Bee                       Parasitic Wasp
##                                667                                  285
##                Buff Tailed Bumblebee                  Carniolan Honey Bee
##                                183                                  152
##                          Bumble Bee                     Italian Honeybee
##                                140                                  113
##                     Japanese Beetle                   Asian Lady Beetle
##                                 94                                   76
##                      Euonymus Scale                            Wireworm
##                                 75                                   69
##                   European Dark Bee                   Minute Pirate Bug
##                                 66                                   62
##                  Asian Citrus Psyllid                      Parastic Wasp
##                                 60                                   58
##               Colorado Potato Beetle                    Parasitoid Wasp
##                                 57                                   51
##                  Erythrina Gall Wasp                      Beetle Order
##                                 49                                   47
##          Snout Beetle Family, Weevil            Sevenspotted Lady Beetle
##                                 47                                   46
##                       True Bug Order                Buff-tailed Bumblebee
##                                 45                                   39
##                        Aphid Family                      Cabbage Looper
##                                 38                                   38
##                  Sweetpotato Whitefly                      Braconid Wasp
##                                 37                                   33
##                        Cotton Aphid                      Predatory Mite
##                                 33                                   33
##                Ladybird Beetle Family                         Parasitoid
##                                 30                                   30
##                        Scarab Beetle                      Spring Tiphia
##                                 29                                   29
##                         Thrip Order                 Ground Beetle Family
##                                 29                                   27
##                   Rove Beetle Family                      Tobacco Aphid
##                                 27                                   27
##                        Chalcid Wasp                Convergent Lady Beetle
##                                 25                                   25
##                       Stingless Bee                   Spider/Mite Class
##                                 25                                   24
```

```
##                Tobacco Flea Beetle                Citrus Leafminer
##                                 24                               23
##                   Ladybird Beetle                    Mason Bee
##                                 23                               22
##                           Mosquito                  Argentine Ant
##                                 22                               21
##                             Beetle       Flatheaded Appletree Borer
##                                 21                               20
##               Horned Oak Gall Wasp              Leaf Beetle Family
##                                 20                               20
##                  Potato Leafhopper       Tooth-necked Fungus Beetle
##                                 20                               20
##                       Codling Moth        Black-spotted Lady Beetle
##                                 19                               18
##                       Calico Scale              Fairyfly Parasitoid
##                                 18                               18
##                        Lady Beetle          Minute Parasitic Wasps
##                                 18                               18
##                          Mirid Bug              Mulberry Pyralid
##                                 18                               18
##                          Silkworm                 Vedalia Beetle
##                                 18                               18
##              Araneoid Spider Order                    Bee Order
##                                 17                               17
##                     Egg Parasitoid                   Insect Class
##                                 17                               17
##           Moth And Butterfly Order     Oystershell Scale Parasitoid
##                                 17                               17
## Hemlock Woolly Adelgid Lady Beetle             Hemlock Wooly Adelgid
##                                 16                               16
##                               Mite                   Onion Thrip
##                                 16                               16
##              Western Flower Thrips                  Corn Earworm
##                                 15                               14
##                  Green Peach Aphid                      House Fly
##                                 14                               14
##                           Ox Beetle              Red Scale Parasite
##                                 14                               14
##                 Spined Soldier Bug           Armoured Scale Family
##                                 14                               13
##                  Diamondback Moth                  Eulophid Wasp
##                                 13                               13
##                  Monarch Butterfly                  Predatory Bug
##                                 13                               13
##              Yellow Fever Mosquito            Braconid Parasitoid
##                                 13                               12
##                       Common Thrip   Eastern Subterranean Termite
##                                 12                               12
##                             Jassid                     Mite Order
##                                 12                               12
##                          Pea Aphid                Pond Wolf Spider
##                                 12                               12
##           Spotless Ladybird Beetle          Glasshouse Potato Wasp
##                                 11                               10
```

```
##                          Lacewing            Southern House Mosquito
##                                10                                 10
##          Two Spotted Lady Beetle                         Ant Family
##                                10                                  9
##                      Apple Maggot                           (Other)
##                                 9                                670
```

Answer: The six most commonly studied species in this dataset are honey bees, parasitic wasps, buff tailed bumblebees, carniolan honey bees, bumble bees, and Italian honeybees all of which are important pollinators for food crops and other agricultural products. Due to their important role in food production, the effect of certain neonicotinoids on these species would likely be of interest to ensure these non-target species are not killed by insecticides.

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

```
# 8. Determining the class of column 'Conc.1..Author.' in the Neonics dataset

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

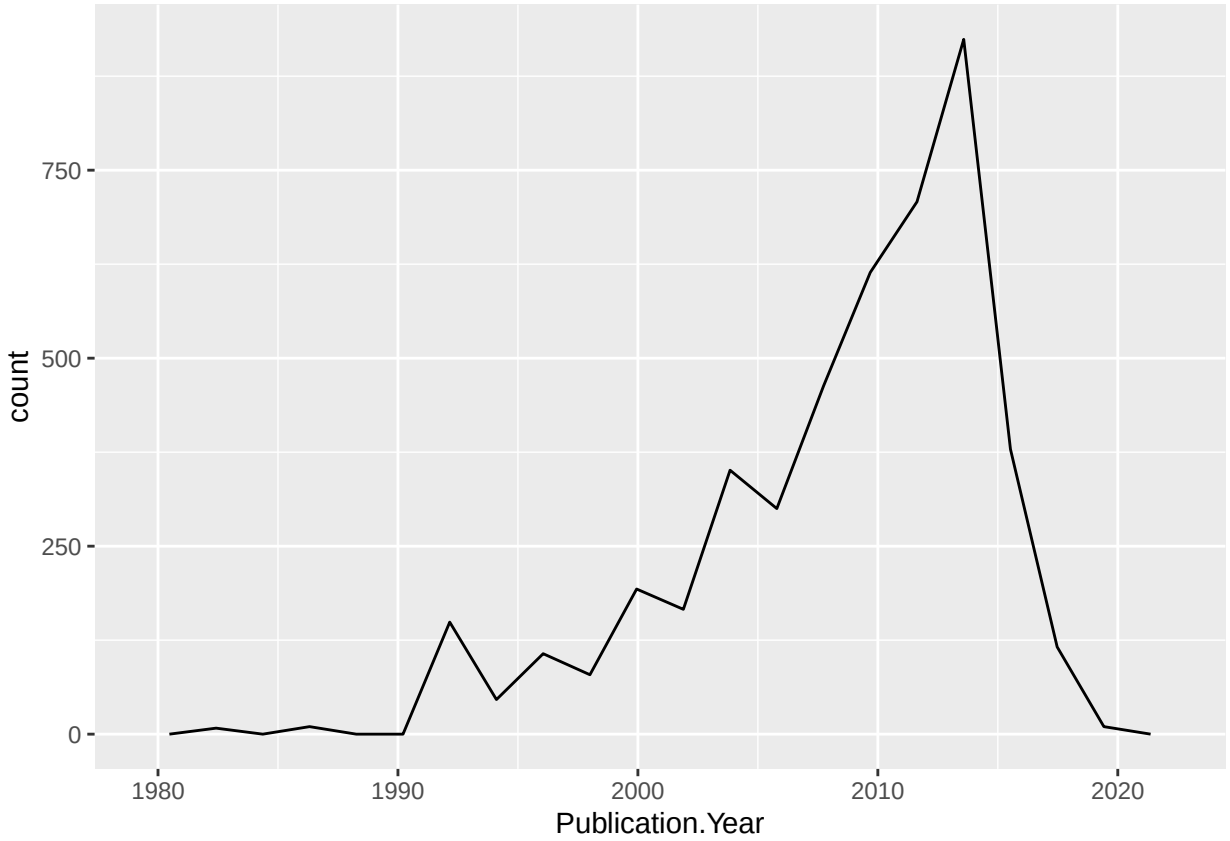
```
## [1] "factor"
```

Answer: The class of columnn "Conc.1..Author." is factor, not numeric. That is because of the subcommand we included when we uploaded and named the Neonics dataset in question #1 to read strings in as factors - "stringsAsFactors = TRUE". However, the values within this column are still numeric.

## Explore your data graphically (Neonics)

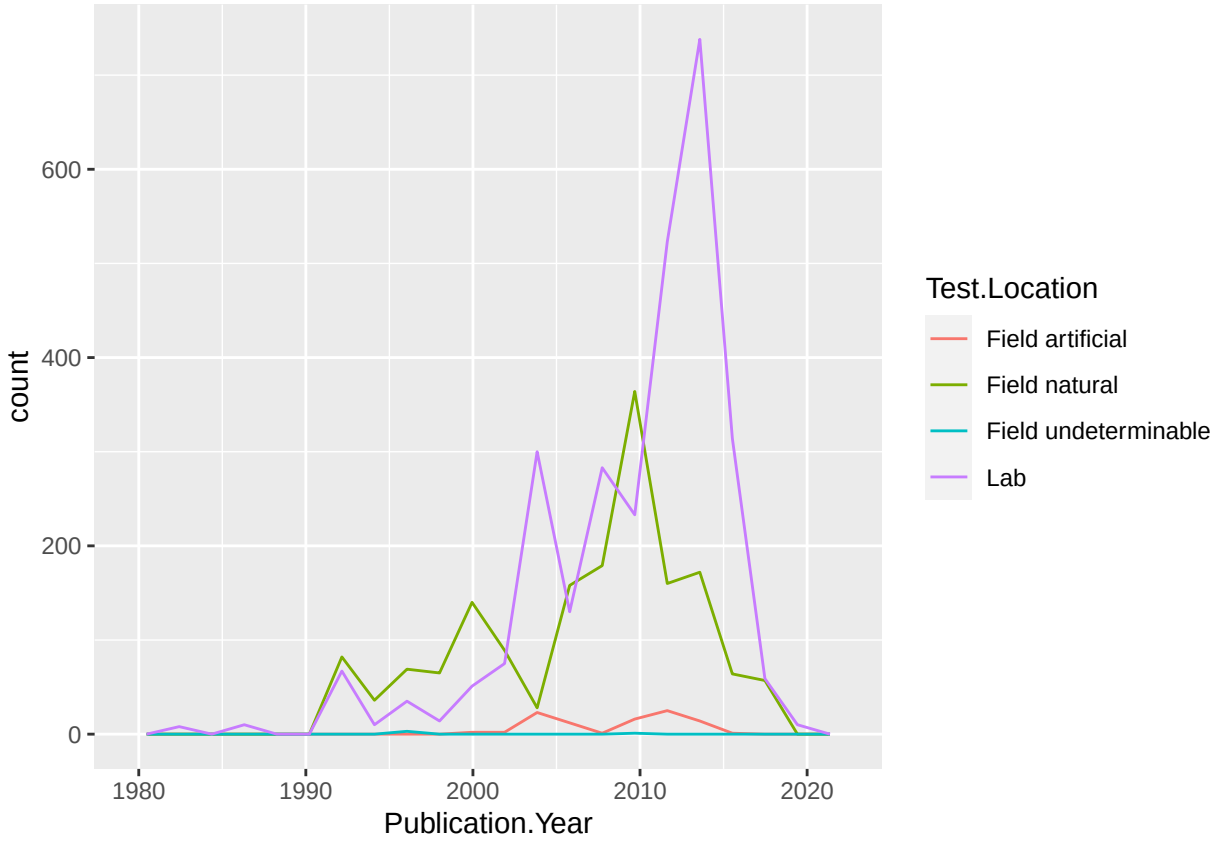
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
# 9. Generating a frequency line graph of the number of studies conducted by
# publication year

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




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

```
# 10. Adding to the frequency line graph in #9, include Test.Location as
# different colors

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

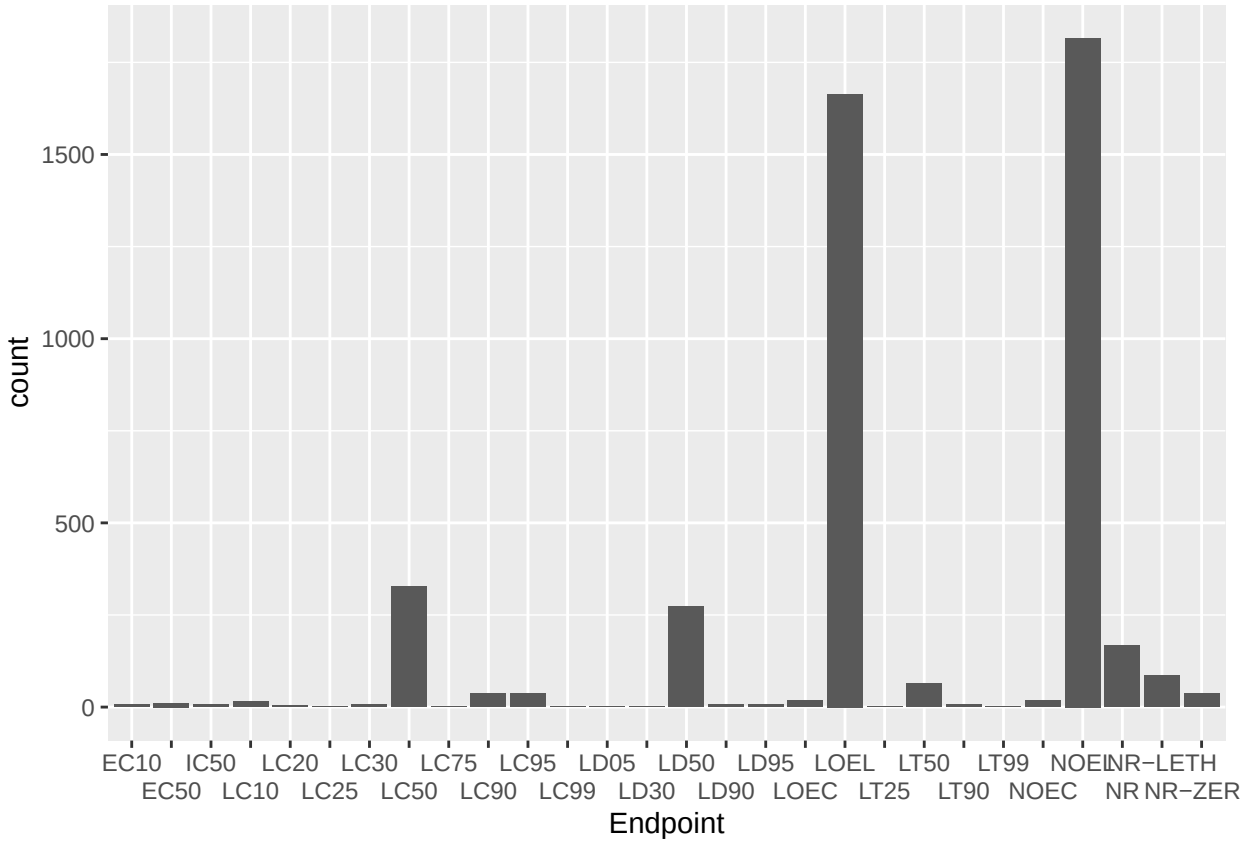



Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are "Labs" and "Field natural," both of which increase greatly from 2000 to 2020. From 1980 to 2000, there were fewer testing locations in total compared to years 2000 to 2020.

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.

```
# 11. Creating a bar graph of Endpoint counts

ggplot(Neonics) + geom_bar(aes(x = Endpoint)) + scale_x_discrete(guide = guide_axis(n.dodge = 2))
```




Answer: LOEL and NOEL are the most common endpoints. LOEL stands for "lowest observable effect level" and describes the lowest dose, in terms of concentration, that produces effects significantly different from responses of controls. NOEL stands for "no observable effect level" and describes the highest dose, in terms of concentration, that produces effects not significantly different from responses of controls.

## 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.

```
# 12. Determining the class of 'collectDate'
class(Litter$collectDate)
```

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

```
# 'collectDate' is a factor, so changing it to a date and then confirming the
# new class of the variable
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
class(Litter$collectDate)
```

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

```
# Determining which dates litter was sampled in August 2018 using the 'unique'
# function
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`?

```
# 13. Running the 'summary' and 'unique' functions to compare them
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
```

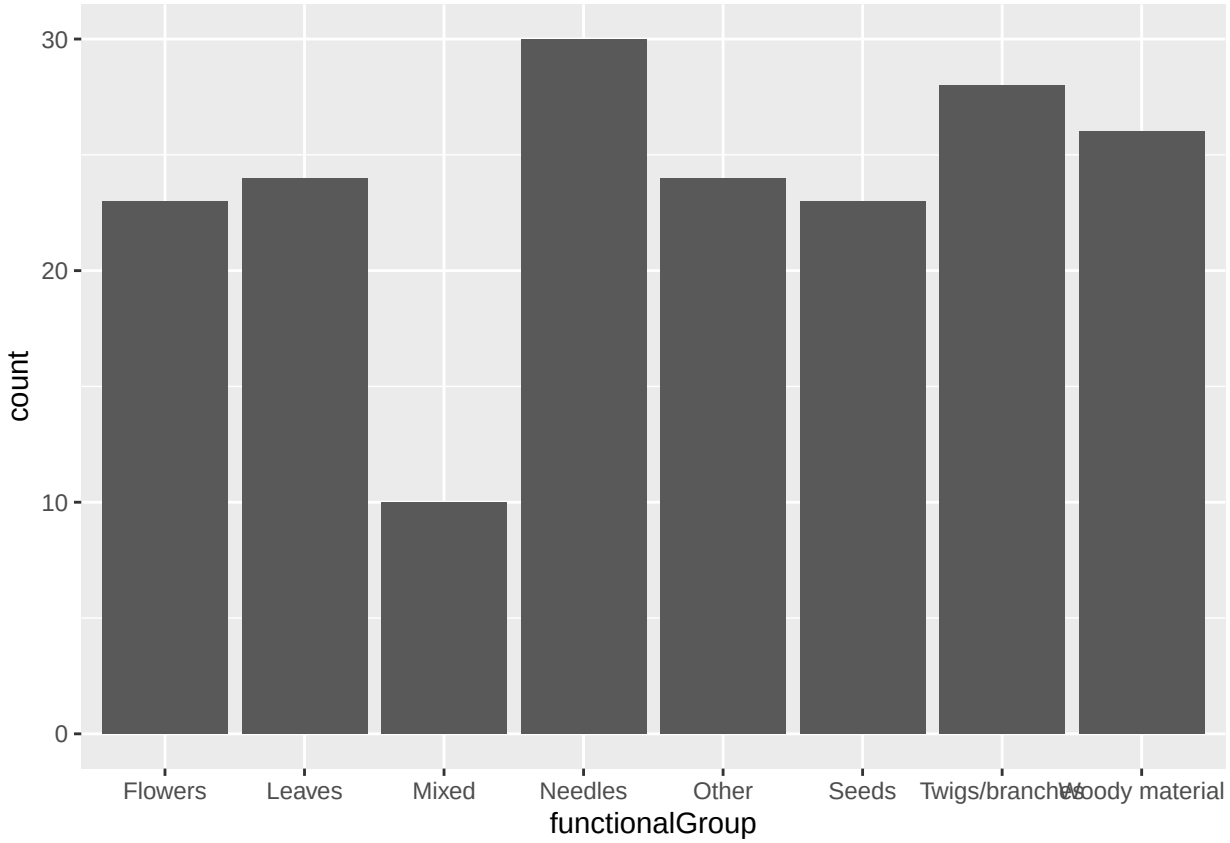
```
unique(Litter$plotID)
```

```
##  [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
```

Answer: In determining how many plots were sampled at Niwot Ridge, the 'summary' function gives you a count of each entry, or factor, that appears in column "plotID." Conversely, the 'unique' function provides a list of all entries that appear in the "plotID" column, excluding duplicates.

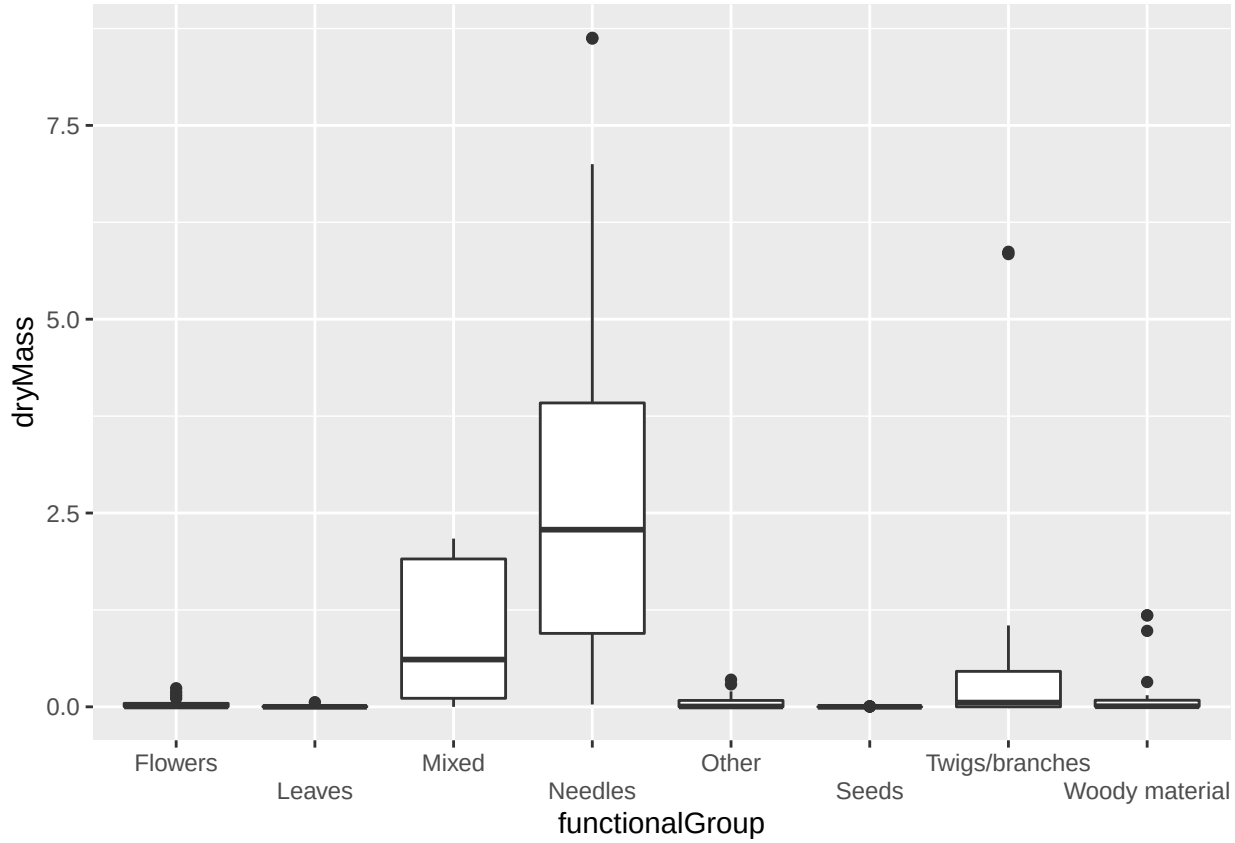
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
# 14. Creating a bar graph of 'functionalGroup' counts

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

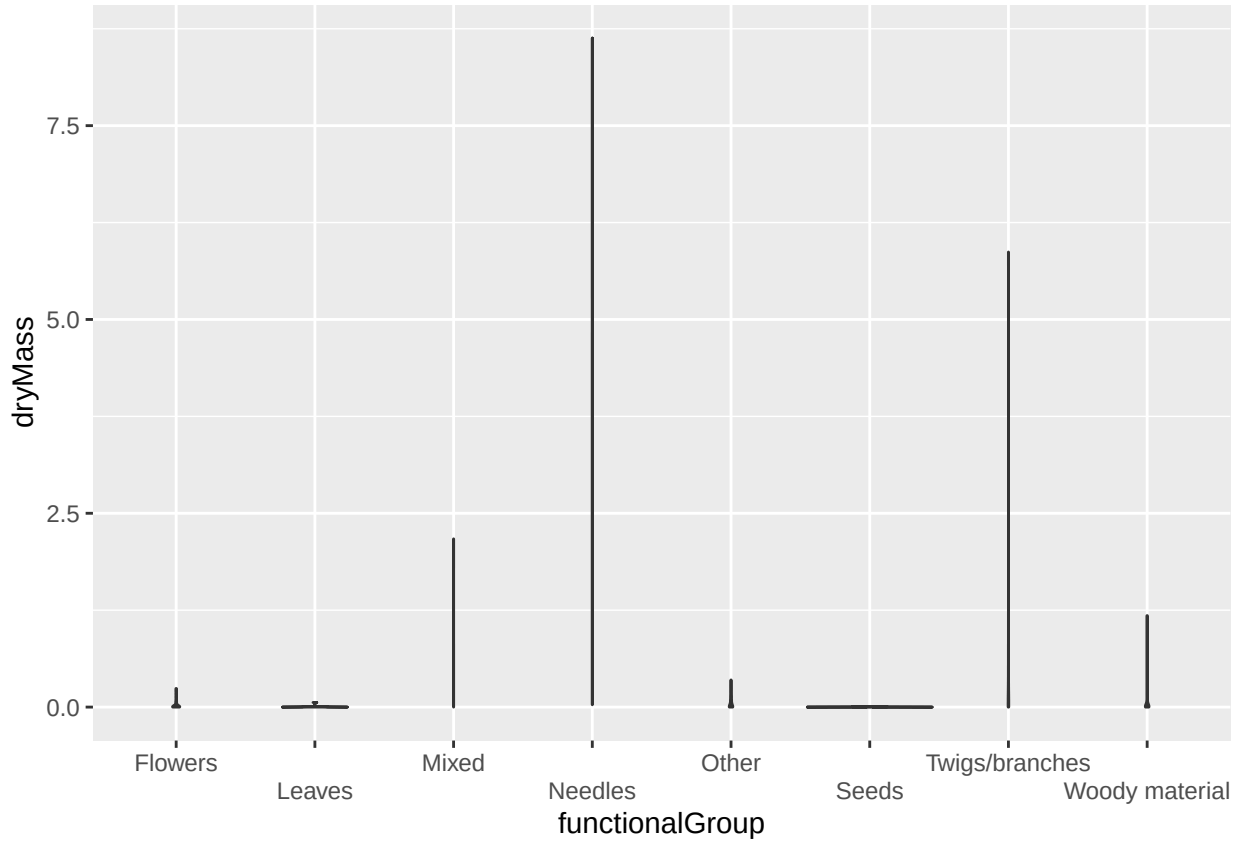



15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
# 15. Creating a boxplot of dryMass by functionalGroup

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




```
# 15. Creating a violin plot of dryMass by functionalGroup

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




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

> Answer: The boxplot above is a more effective visualization option than the violin plot because of the spread, or distribution, of the data. Since there are not a lot of data points, or counts, clustered around the mean or interquartile range values, the lack of width across the violin plot reflects that. As a result, the boxplots allows for a better visualization of the distribution of the data for each functional group.

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

> Answer: Needles, Mixed, and Twigs/Branches tend to have the highest biomass at these sites, as shown in the boxplot above.