# Assignment 3: Data Exploration

Sammy DiLoreto

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.

```
#Check working directory
getwd()
```

```
## [1] "/Users/sammydiloreto/Library/CloudStorage/Box-Box/ENV872-EDA/EDA-Spring2023/Assignments"
```

```
#Load necessary packages
library(tidyverse)
library(lubridate)
#Upload two data sets
Neonics <- read.csv("../Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = T)
Litter <- read.csv("../Data/Raw/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: Insecticides are used to get rid of insects that pose harm to crops, although some insecticides impact other insects and organisms that are not targeted, this is called off-target effects. Neonicotinoids are less toxic to vertebrates and mammals compared to other pesticides but they have off-target effects on other insects. Understanding the ecotoxicology of neonicotinoids is important to understand how these compounds target insects, cause toxicity, and which insects may be more impacted. This knowledge informs regulation on which plants and insects neonicotinoids can be applied to in order to limit off-target effects. In particular, neonicotinoids are toxic to bees, which are crucial for pollination, so its important to understand how they are affected and how best to limit those impacts.

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 could promote the start and spreading of forest fires. Studying the type of litter and dry mass can indicate fire risk and gives clues about the rate of forest decay. The type of litter present on the forest floor also provides evidence of differences in composition of forest soil. Woody debris is important in carbon budgets and nutrient cycling and it provides a source of energy and habitat for various organisms and ecosystems, therefore understanding the distribution of debris in a forest can indicate types of organisms present.

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. Mass of various functional groups of litter and fine woody debris is measured in this sampling network. These functional groups are leaves, needles, twigs/branches, woody material, seeds, flowers and other non-woody reproductive structures, other, and mixed (unsorted material). 2. Litter is collected from evelated traps and fine woody debris are collected from ground traps. Litter is defined as material that is dropped from the forest canopy and has an end diameter of <2cm and a length <50 cm. Fine woody debris is defined as material that is dropped from the forest canopy and has an end diameter <2cm and a length >50 cm. 3. Sampling is done from tower plots and the locations of these plots are randomly selected within the 90% flux footprint of the primary and secondary airsheds. Placement of traps within the plots was either targeted or randomized, depending on the vegetation. Ground traps are sampled once per year. Sampling frequency for elevated traps varies by vegetation at the site.

## Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#dimensions of Neonics dataset
dim(Neonics)
```

```
## [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?

```
#determine 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 commonly studied effects are mortality and population. Mortality means that the direct cause of death is by the chemical so this important to study to understand which insects are dying from direct use of neonicotinoids and how many. Population refers to population level effects such as abundance which would provide information on the number of individuals of a taxon per unit area equivalent to density. If abundance appears to be lower than what is expected this would give indication of potential chemical effect.

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...]

```
#determine the six most commonly studied species
#and sort descending to see the most common at the top
sort(summary(Neonics$Species.Common.Name), decreasing= TRUE)
```

```
##                          (Other)                       Honey Bee
##                              670                             667
##                    Parasitic Wasp           Buff Tailed Bumblebee
##                              285                             183
##               Carniolan Honey Bee                      Bumble Bee
##                              152                             140
##                  Italian Honeybee                Japanese Beetle
##                              113                              94
##                Asian Lady Beetle                 Euonymus Scale
##                               76                              75
##                         Wireworm              European Dark Bee
##                               69                              66
##                Minute Pirate Bug              Asian Citrus Psyllid
```

```
##                          62                          60
##               Parastic Wasp        Colorado Potato Beetle
##                          58                          57
##             Parasitoid Wasp          Erythrina Gall Wasp
##                          51                          49
##              Beetle Order   Snout Beetle Family, Weevil
##                          47                          47
##      Sevenspotted Lady Beetle              True Bug Order
##                          46                          45
##         Buff-tailed Bumblebee               Aphid Family
##                          39                          38
##              Cabbage Looper        Sweetpotato Whitefly
##                          38                          37
##               Braconid Wasp              Cotton Aphid
##                          33                          33
##              Predatory Mite        Ladybird Beetle Family
##                          33                          30
##                 Parasitoid              Scarab Beetle
##                          30                          29
##              Spring Tiphia               Thrip Order
##                          29                          29
##         Ground Beetle Family         Rove Beetle Family
##                          27                          27
##              Tobacco Aphid              Chalcid Wasp
##                          27                          25
##        Convergent Lady Beetle              Stingless Bee
##                          25                          25
##           Spider/Mite Class         Tobacco Flea Beetle
##                          24                          24
##            Citrus Leafminer             Ladybird Beetle
##                          23                          23
##                  Mason Bee                    Mosquito
##                          22                          22
##               Argentine Ant                      Beetle
##                          21                          21
##    Flatheaded Appletree Borer        Horned Oak Gall Wasp
##                          20                          20
##           Leaf Beetle Family          Potato Leafhopper
##                          20                          20
##    Tooth-necked Fungus Beetle              Codling Moth
##                          20                          19
##     Black-spotted Lady Beetle              Calico Scale
##                          18                          18
##          Fairyfly Parasitoid               Lady Beetle
##                          18                          18
##        Minute Parasitic Wasps                 Mirid Bug
##                          18                          18
##            Mulberry Pyralid                  Silkworm
##                          18                          18
##              Vedalia Beetle        Araneoid Spider Order
##                          18                          17
##                  Bee Order              Egg Parasitoid
##                          17                          17
##               Insect Class     Moth And Butterfly Order
```

```
##                                17                                17
##       Oystershell Scale Parasitoid Hemlock Woolly Adelgid Lady Beetle
##                                17                                16
##              Hemlock Wooly Adelgid                              Mite
##                                16                                16
##                        Onion Thrip              Western Flower Thrips
##                                16                                15
##                       Corn Earworm                  Green Peach Aphid
##                                14                                14
##                          House Fly                         Ox Beetle
##                                14                                14
##                 Red Scale Parasite                 Spined Soldier Bug
##                                14                                14
##              Armoured Scale Family                   Diamondback Moth
##                                13                                13
##                       Eulophid Wasp                 Monarch Butterfly
##                                13                                13
##                       Predatory Bug              Yellow Fever Mosquito
##                                13                                13
##                 Braconid Parasitoid                       Common Thrip
##                                12                                12
##       Eastern Subterranean Termite                             Jassid
##                                12                                12
##                           Mite Order                        Pea Aphid
##                                12                                12
##                   Pond Wolf Spider           Spotless Ladybird Beetle
##                                12                                11
##             Glasshouse Potato Wasp                          Lacewing
##                                10                                10
##            Southern House Mosquito            Two Spotted Lady Beetle
##                                10                                10
##                          Ant Family                       Apple Maggot
##                                 9                                 9
```

Answer: Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee are the six most commonly studied species in this dataset, aside from "other." These are the most commonly studies because neonicotinoids are toxic to bees. Neonicotinoids are absobed into the plant and present in pollen, thus exposing 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?

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

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

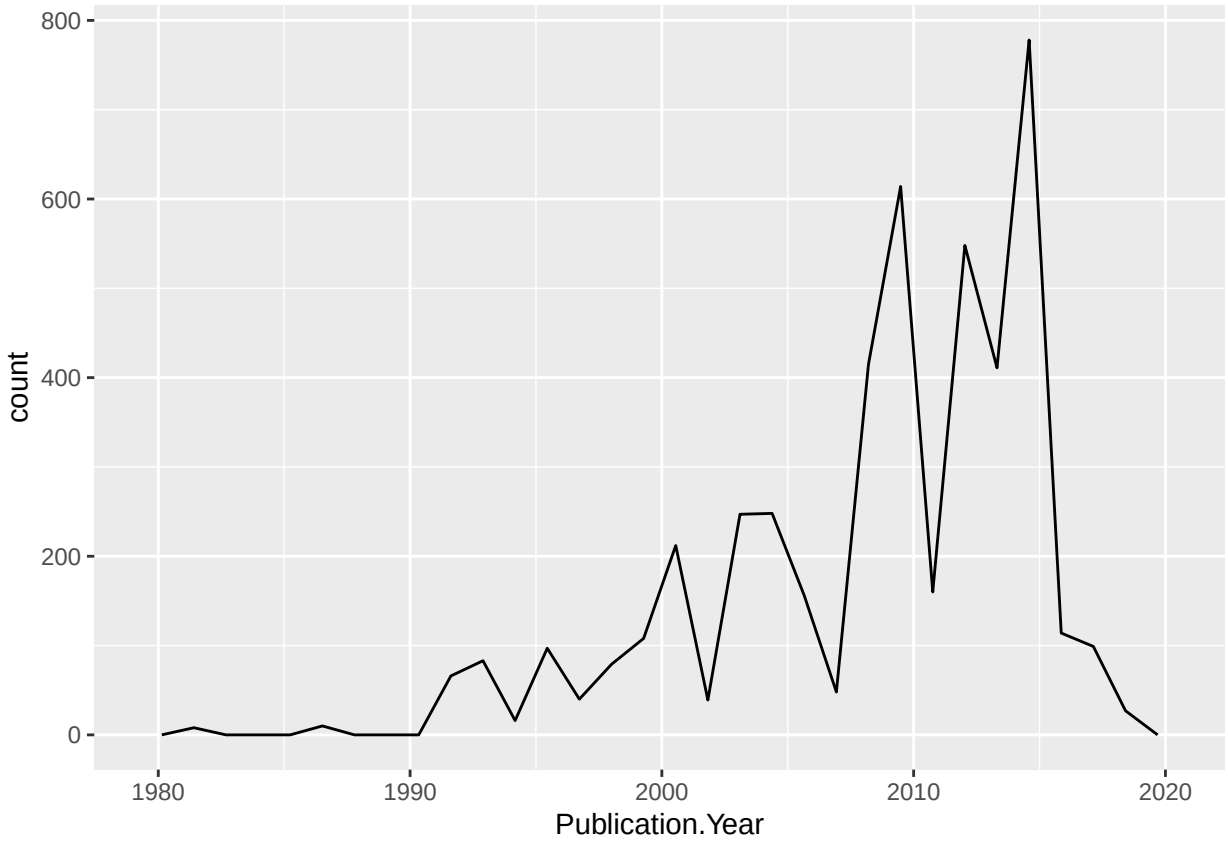
Answer: The class is factor. It is not numeric because the concentrations have different units therefore you cannot compare them. If they were numeric, you would be able to do summary statistics for them but that would not make sense because each value means something different based on its units.

## Explore your data graphically (Neonics)

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

```
#plot of number of studies conducted by publication year
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year))
```

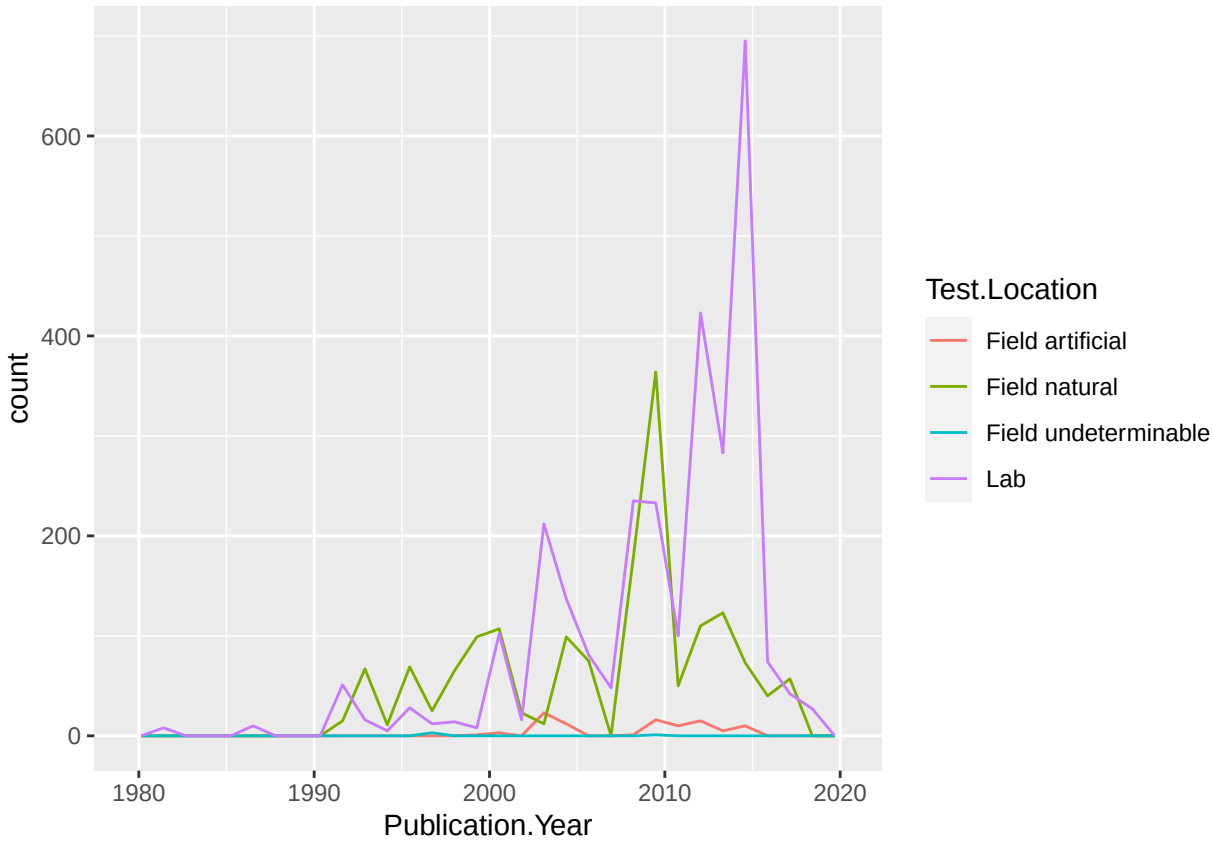
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




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

```
#plot of number of studies conducted by publication year, sorted by Test location
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location))
```

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



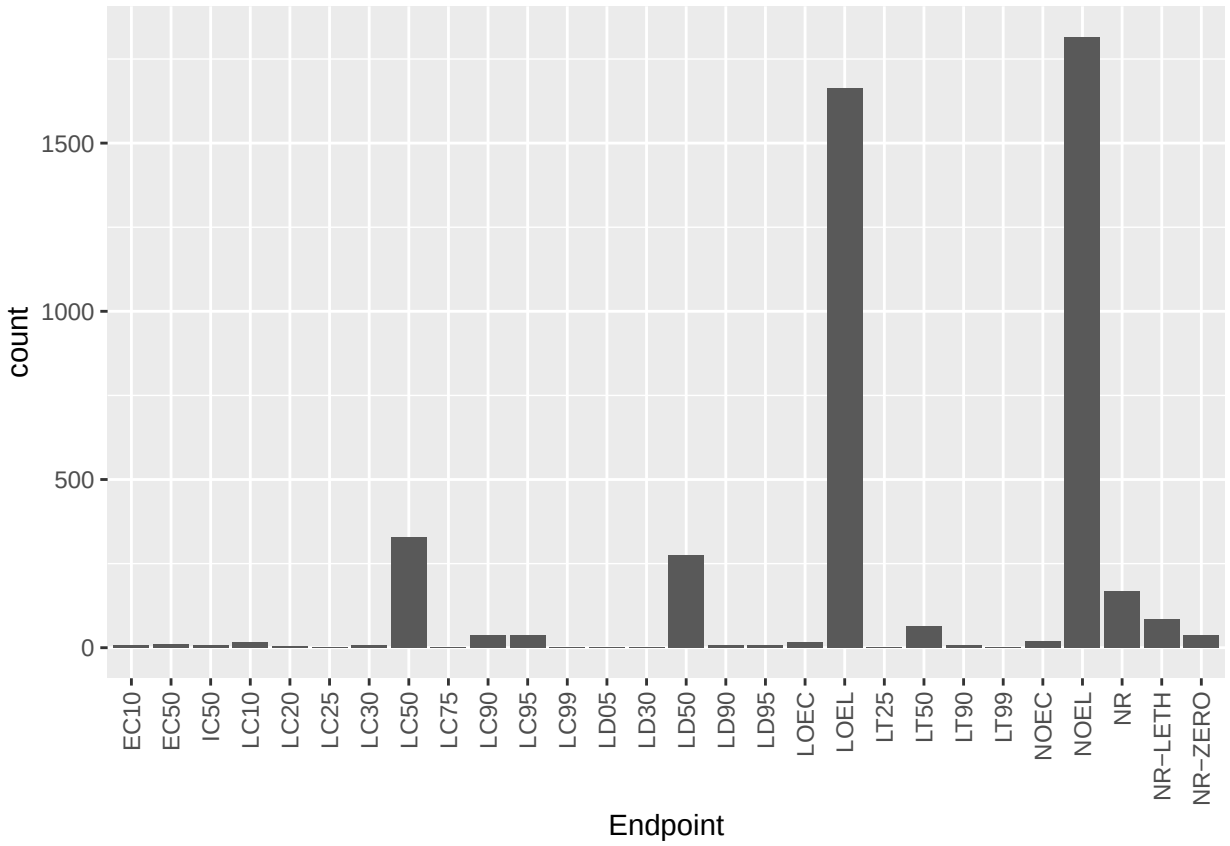

Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: Lab and field natural are the most common test locations. In the 2010s, field natural was more popular but in more recent years, lab based studies have increased. Field artifical and field underterminable have never been super common.

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...]

```
#bar graph of endpoint counts
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar()+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```




Answer: LOEL and NOEL are the two most common endpoints. LOEL or lowest observed effect level is the lowest dose that causes effects that were significantly different from the responses of controls. NOEL or no observed effect level is the highest dose at which there is not significantly different results from the responses of controls according to statistical analyses.

## 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.

```
#Determine the class of collectDate
class(Litter$collectDate)
```

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

```
#Change collectDate to date format
Litter$collectDate <- ymd(Litter$collectDate)
#Confirming the new class of collectDate
class(Litter$collectDate)
```

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

```
#Determine which dates litter was sampled in August 2018
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`?

```
#number of plots at Niwot Ridge
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
```

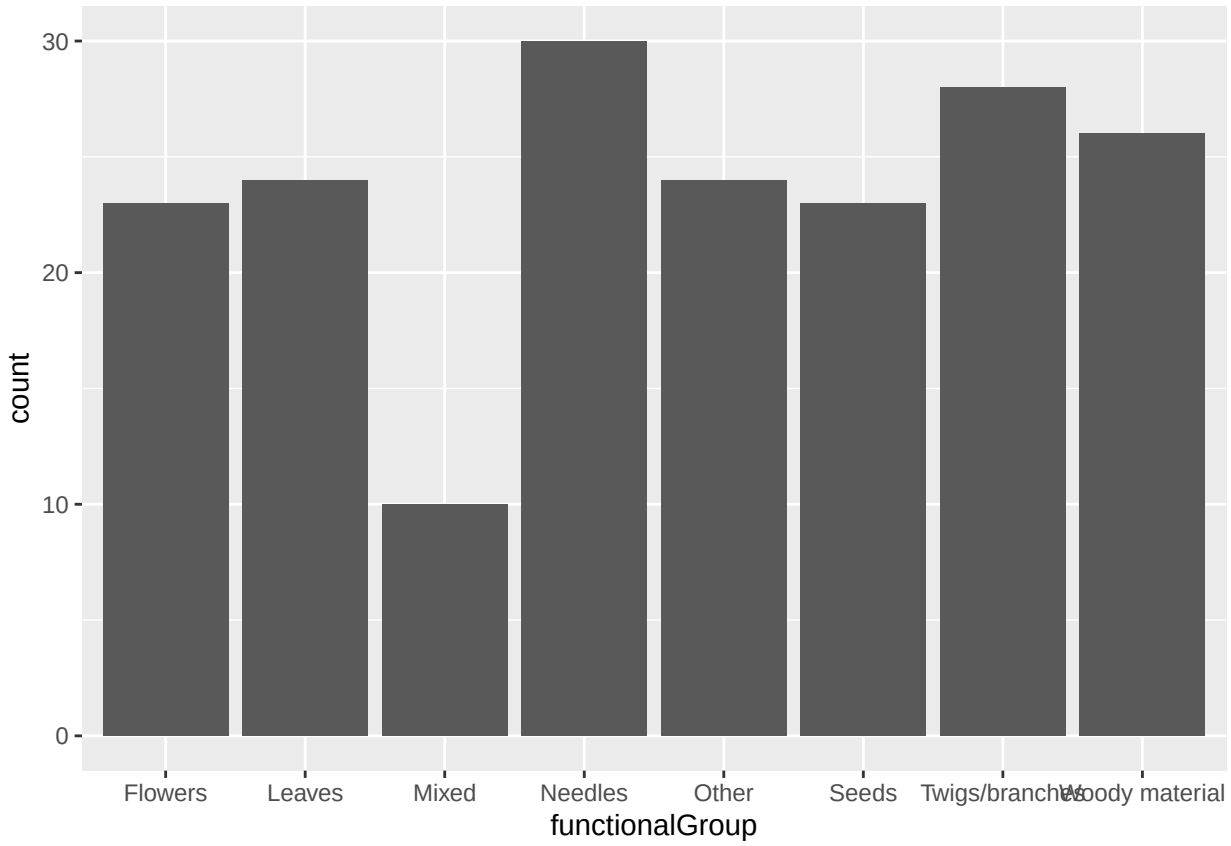
```
summary(Litter$plotID) #testing what summary shows
```

```
## 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 unique plots at Niwot Ridge. The unique function will pull out all the unique values and specify how many unique values. The summary function will also pull out all the unique values but it will speciify how many of each unique value are in the column.

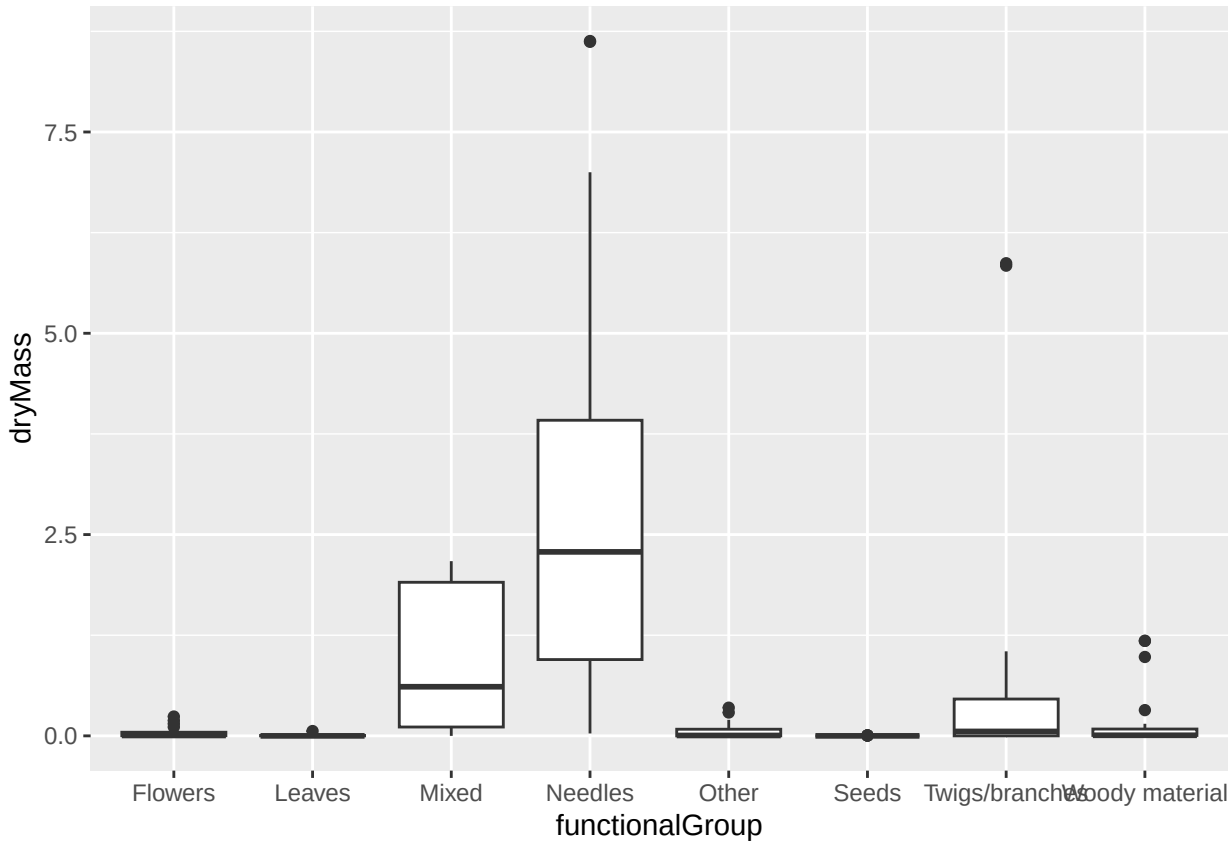
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#bar graph of functionalGroup counts
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar()
```

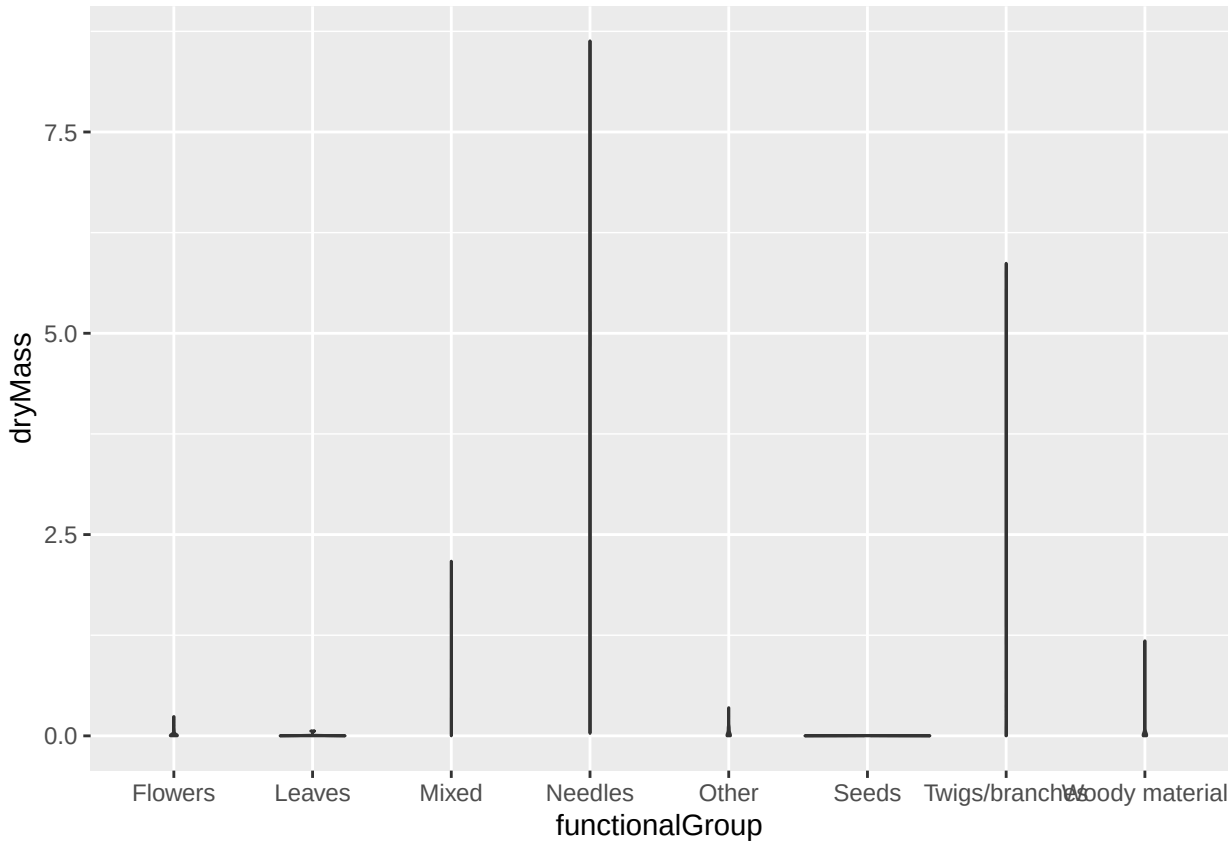



15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




```
#violin plot of dryMass by functionalGroup
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass))
```




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

> Answer: Violin plots are used to show the distribution within the data rather than just the summary statistics like a boxplot. If there is not a lot of distribution within the summary statistics of data, like in mixed, needles, twigs/branches, and woody material, a violin plot will just show up as a vertical line. Or if all data is around the same value like for leaves and seeds, the violin plot will be a horizontal line because all the distribution is around one value. Also, there is a lot of variability in dry mass between functional groups which makes it difficult to view the distributions next to each other because the y axis is very broad. Therefore, a boxplot is a more effective visualization of this data to gain an understanding of where the highest biomass is at these sites.

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

> Answer: Needles and mixed litter tend to have the highest biomass. This is seen on the boxplot rather than the violin plot because the violin plot includes the outliers so it appears that needles and twigs/branches have the highest biomass but that is only because twigs/branches have a very large outlier. Whereas on the boxplot, needles and mixed litter have the highest medians.