# Assignment 3: Data Exploration

Isaac Benaka

## 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.

```
library(tidyverse)
library(lubridate)
setwd("~/Desktop/Fall 2022/872 - Data Analytics/EDA-Fall2022")
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: We are interested in neonictonoid impacts on insects because the pesticides may be impacting other insect species than the ones they are targeting on crops. Neonictonoids may also trickle into the wider food web by predators eating the insects they act on.

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: We might be interested in studying woody debris data because forest litter that falls onto the ground contributes to decomposition and nutrient cycling.

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. 0.5 m^2 mesh traps, elevated 80 cm above the ground, were used to collect litter and woody desbris. 2. In forested tower airsheds, sampling took place over 20 40m x 40m plots. 3. Ground traps were sampled once per year.

## Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the 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?

```
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: Based on the summary, mortality and population were the most common effects that were studied. These effects were likely studied the most because they are the most important factors when it comes to population modeling.

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)
```

```
##                          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 are Honey bee, Parasitic wasp, Buff tailed bumblebee, Carniolan honey bee, Bumble bee, and Italian honey bee. All these species are key pollinators. Because they provide valuable ecosystem services, they may be of interest over other insects.

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"
```

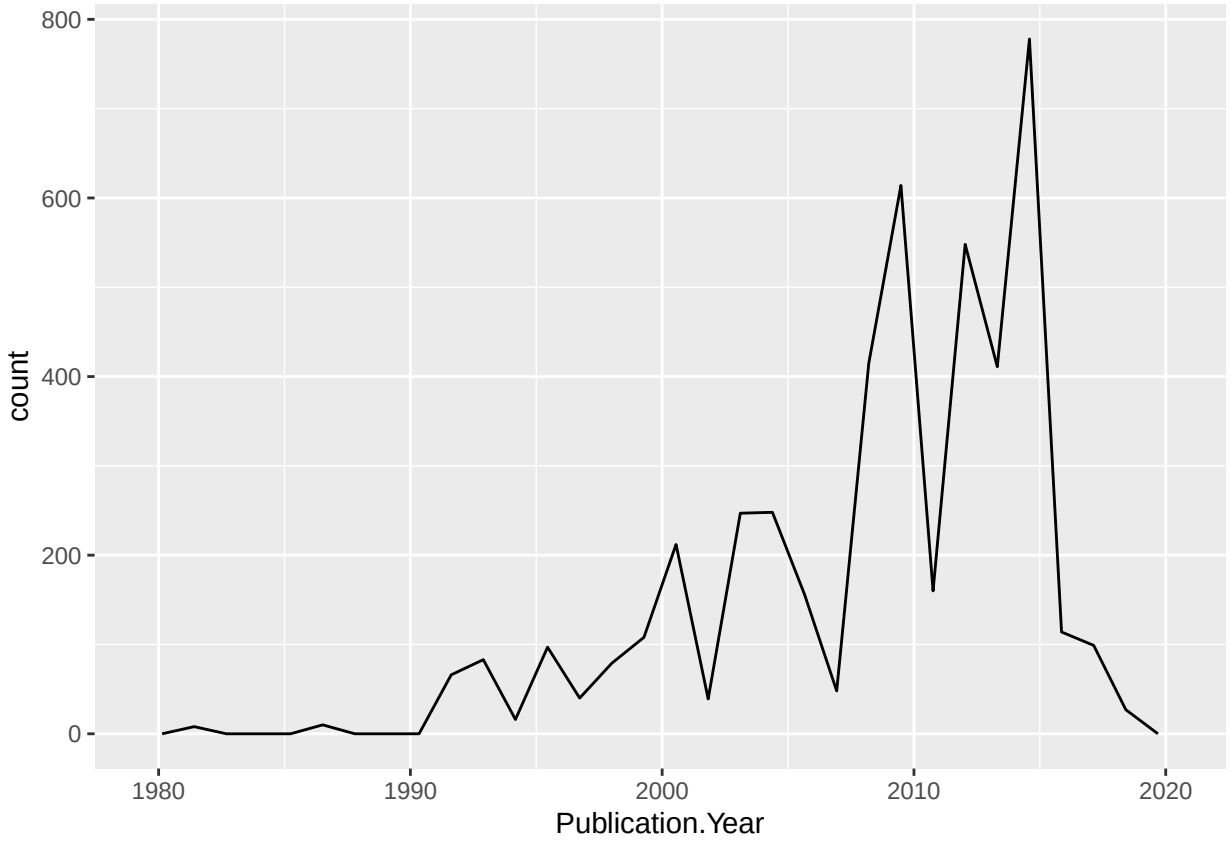
Answer: The class of Conc.1..Author. is a factor. This class is a factor instead of numeric because these are categorical values assigned to each author.

## 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`.
```



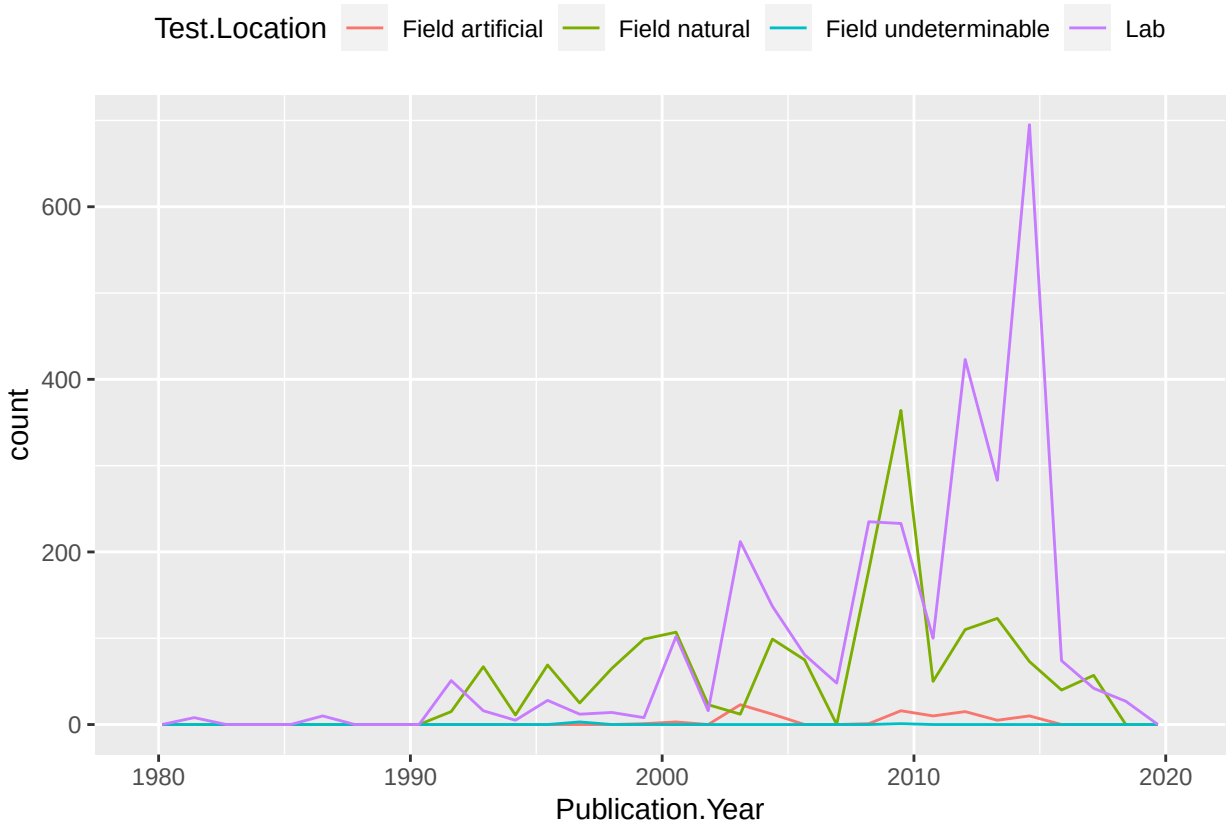

10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
class(Neonics$Test.Location)
```

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

```
ggplot(Neonics) +
  geom_freqpoly(aes(color = Test.Location, x = Publication.Year)) +
  theme(legend.position = "top")
```

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

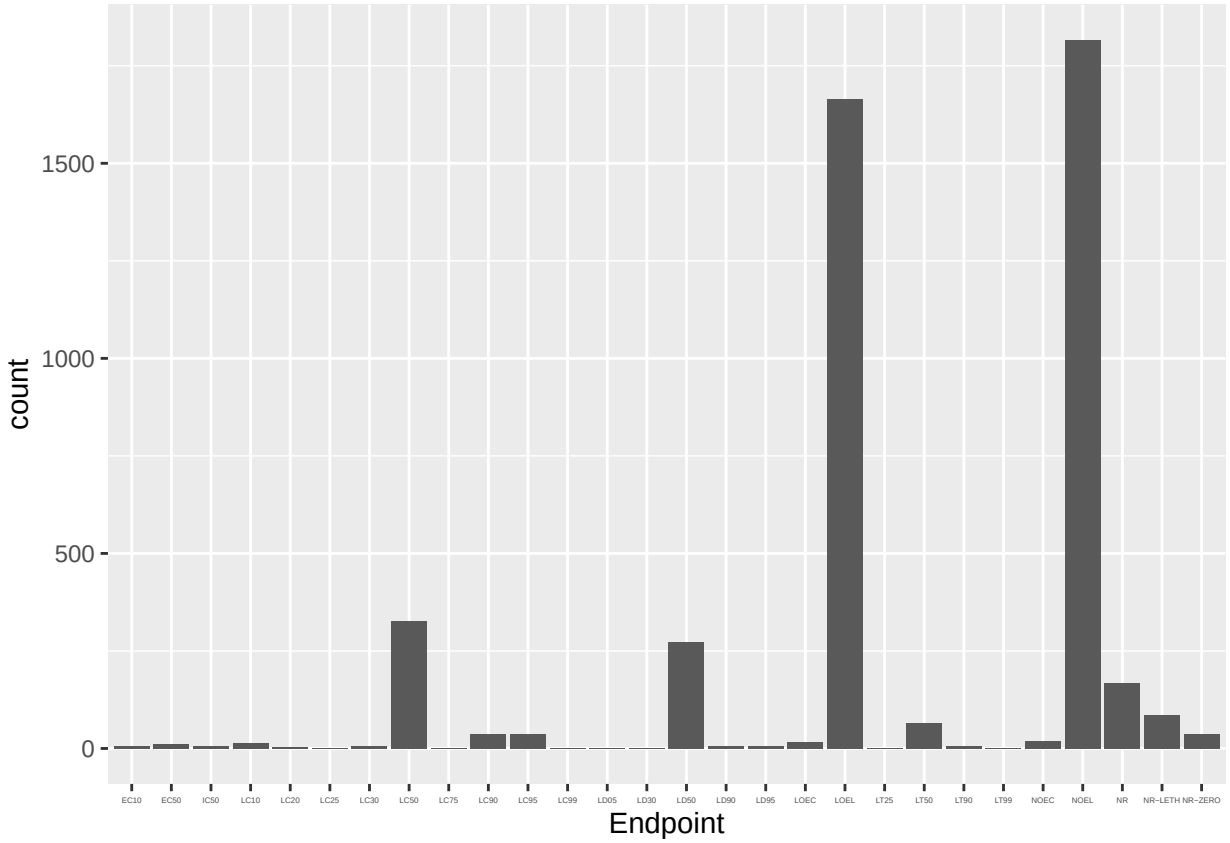



Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are "Lab" and "Field natural." Starting in 2010, "Lab" test locations increased as "Field natural" locations decreased.

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.

```
Endpointplots <- ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar()
Endpointplots + theme(axis.text.x = element_text(size = 3))
```




Answer: LOEL and NOEL are the two most common end points. LOEL stands for lowest-observable-effect-level, meaning that was the lowest dose of chemicals that produced effects that were significantly different than the control samples. NOEL stands for no-observable-effect-level, which means that is the highest dose of chemical that does not produce a significantly different response from the control sample.

## 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) #This class was originally a factor.
```

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

```
Litter$collectDate <- ymd(Litter$collectDate)
class(Litter$collectDate) #Recheck the class of collectDate to view new class.
```

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

```
unique(Litter$collectDate) #Litter was sampled on the 2nd and the 30th in August 2018.
```

```
## [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`?

```
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
```

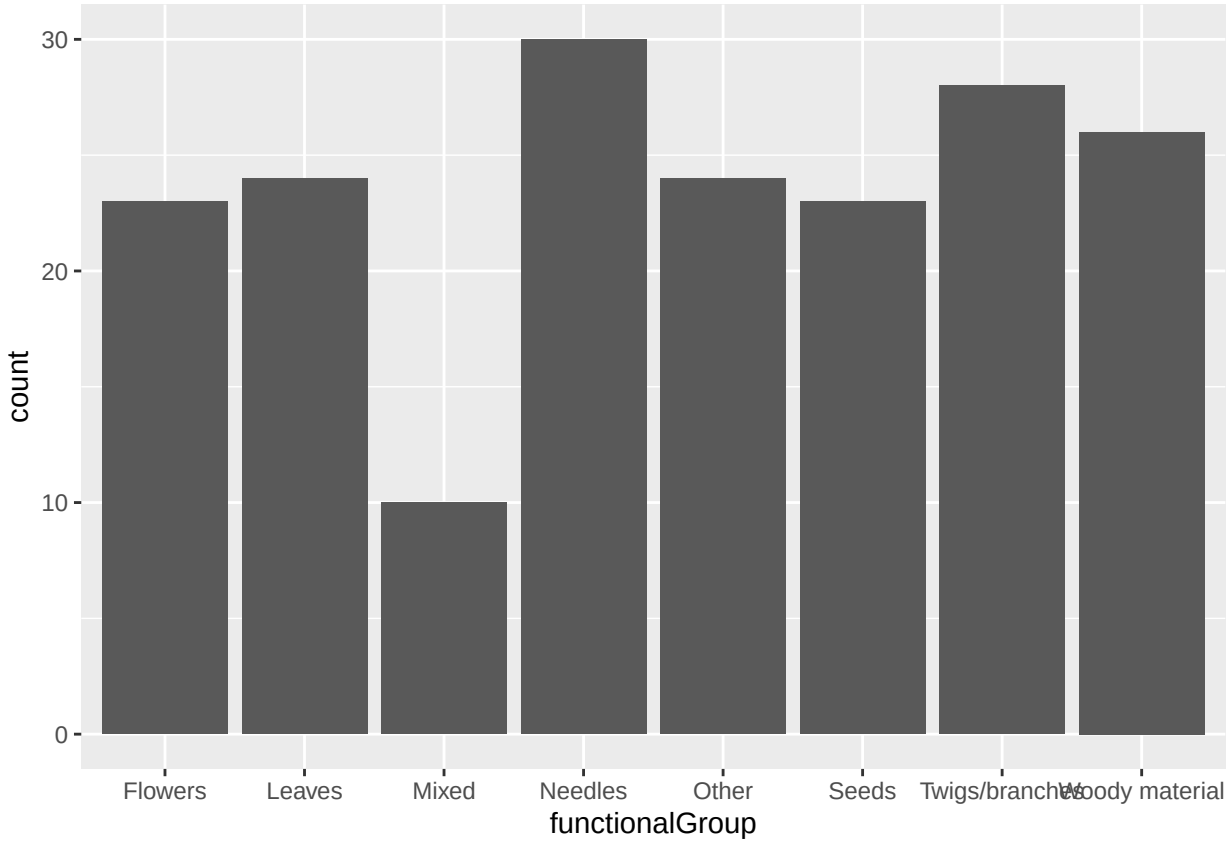
```
summary(Litter$plotID)
```

```
## 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: the 'unique' function only returns the names of the different plots, whereas the 'summary' function returns how many samples correspond with each plot.

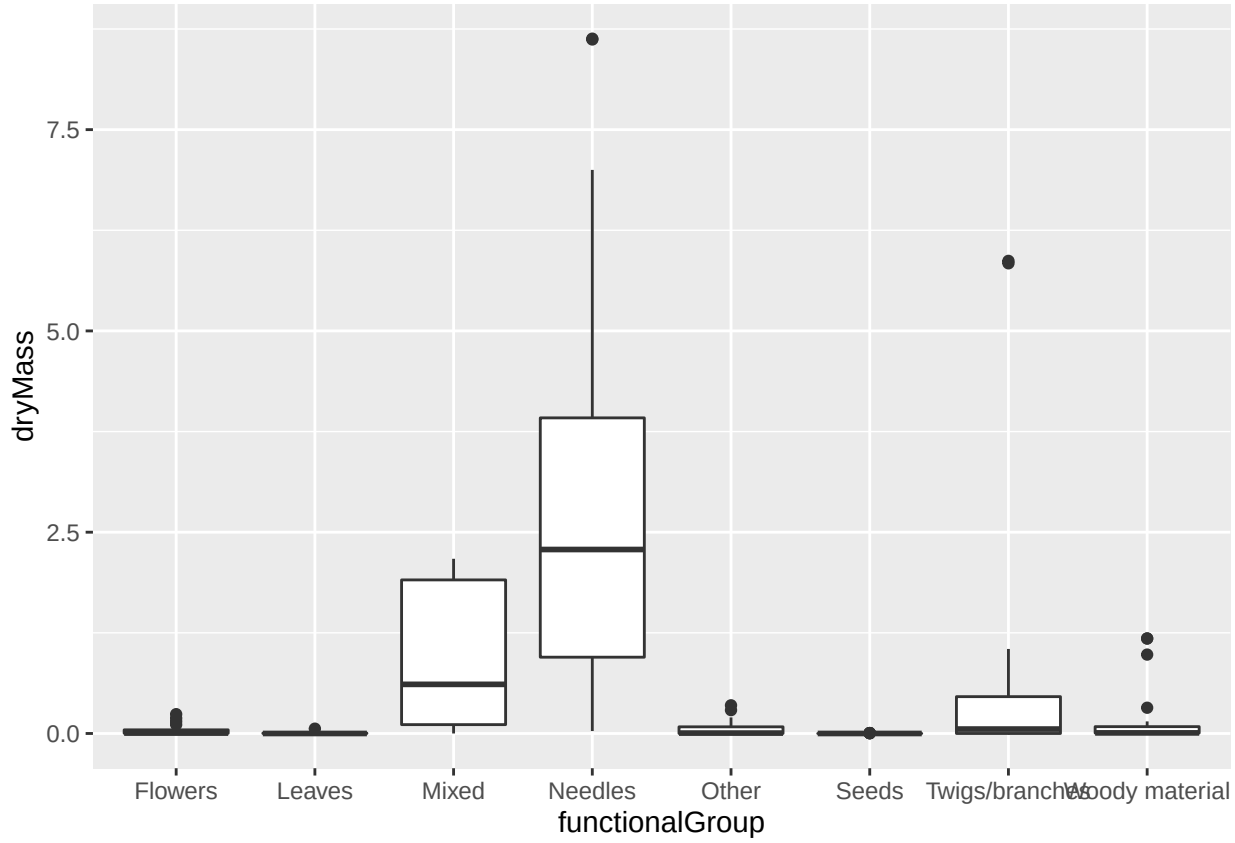
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

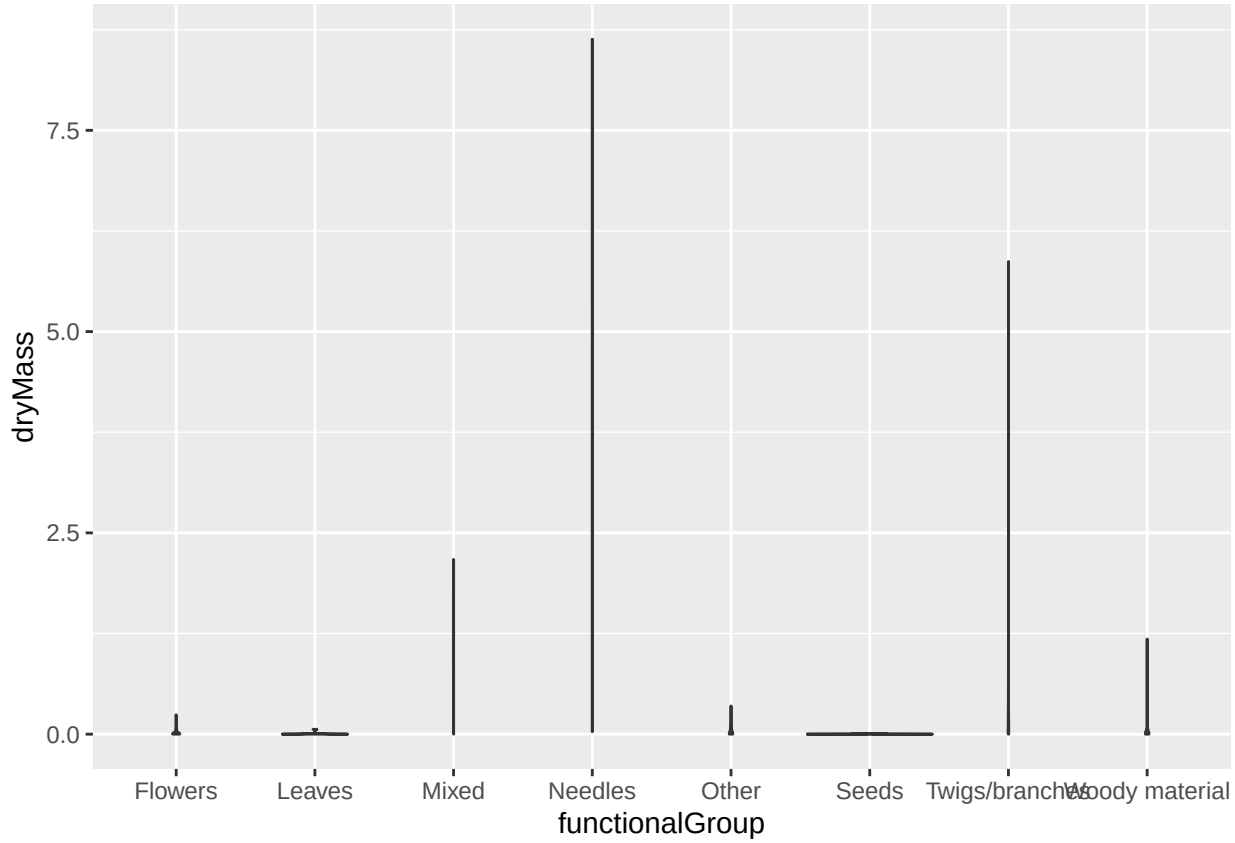


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




```
#Violin plot
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: With the boxplot, we are able to see the interquartile range, mean, and outliers more clearly than with the violin plot.

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

> Answer: Needles, mixed litter, and twigs/branches tended to compose the largest amount of biomass at the sites.