# Assignment 3: Data Exploration

Emma Kaufman

Fall 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.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)

getwd()
```

```
## [1] "/home/guest/EDE_Fall2023"
```

```
#reading in the datasets
#reading in ECOTOX neonicotinoid dataset
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)
#reading in Niwot Ridge NEON dataset for litter and woody debris
```

## 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: Understanding the ecotoxicology of these insecticides can help us to understand the risks neonicotinoids pose to other non-target insects. What impact do they have on important pollinators (that are not the target), or is there an impact on the birds or mammals that eat the target insects?

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 are important for nutrient cycling in forest ecosytems. When this debris breaks down it releases important nutrients into the forest soil. Litter and woody debris also are sources of microbial diversity within forests. That being said, high accumulation of litter and woody debris can be dangerous in areas that are at risk for forest fires, and litter and woody debris can sometimes be classified as "fuels" that need to be reduced. As a result we might be interseted in studying litter and woody debris to understand soil health, microbial diversity, or forest fire risk in a given forest.

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 material ("defined as material that is dropped from the forest canopy and has a butt end diameter <2cm and a length <50 cm") was collected in elevated 0.5 m^2 pvc traps. 2. Sampling occured at terrestrial NEON sites that contained woody vegetation >2m tall 3. Sites were randomly placed ("In sites with > 50% aerial cover of woody vegetation >2m in height, placement of litter traps was random and utilized the randomized list of grid cell locations also used for herbaceous clip harvest and belowground biomass sampling").

## Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
dim(Neonics) #getting dimensions
```

```
## [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?

```r
summary(Neonics$Effect) #summary function
```

```
##     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 and mortality. I think these effects are specifically of interest because scientists want to understand what insects this insecticide is targeting, and whether or not the insecticide is effective at killing and reducing population sizes for pests that negatively impact agriculture.

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...]

```r
summary(Neonics$Species.Common.Name) #summary function
```

```
##                            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: Honey Bee, Parastic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee are the six most commonly studied species in the dataset. These species are all pollinators. They are of interest because of their ecological importance in the reproduction cycle of plants.

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?

```r
class(Neonics$Conc.1..Author.) #what class is it
```

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

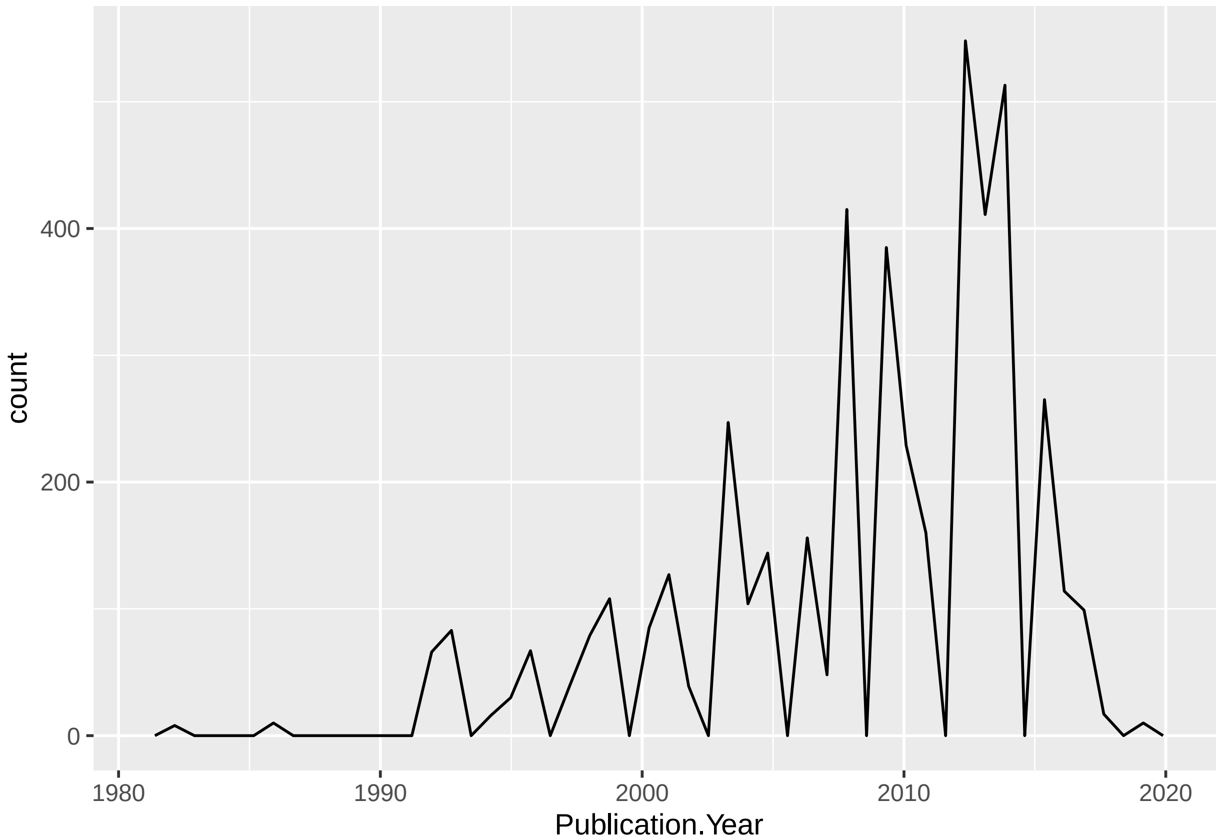
```r
#View(Neonics$Conc.1..Author.) #viewing the column to understand the class
```

Answer: It is a factor class. It is not numeric because it contains categorical data of a threshold concentration. We also read strings in as factors.

## Explore your data graphically (Neonics)

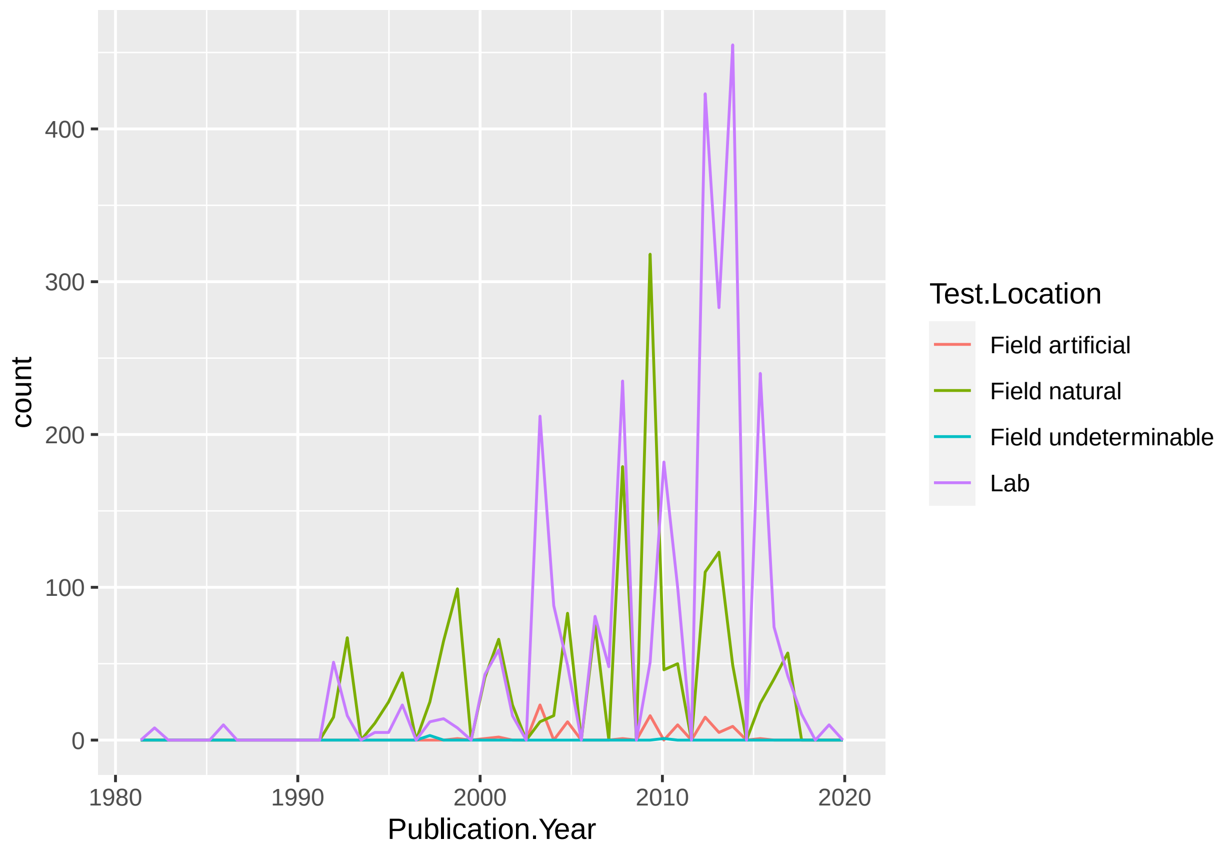
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```r
#plot of number of studies by publication year
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50,  lty = 1)
```




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

```r
#adding color aesthetic for different test locations
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 50, lty = 1)
```



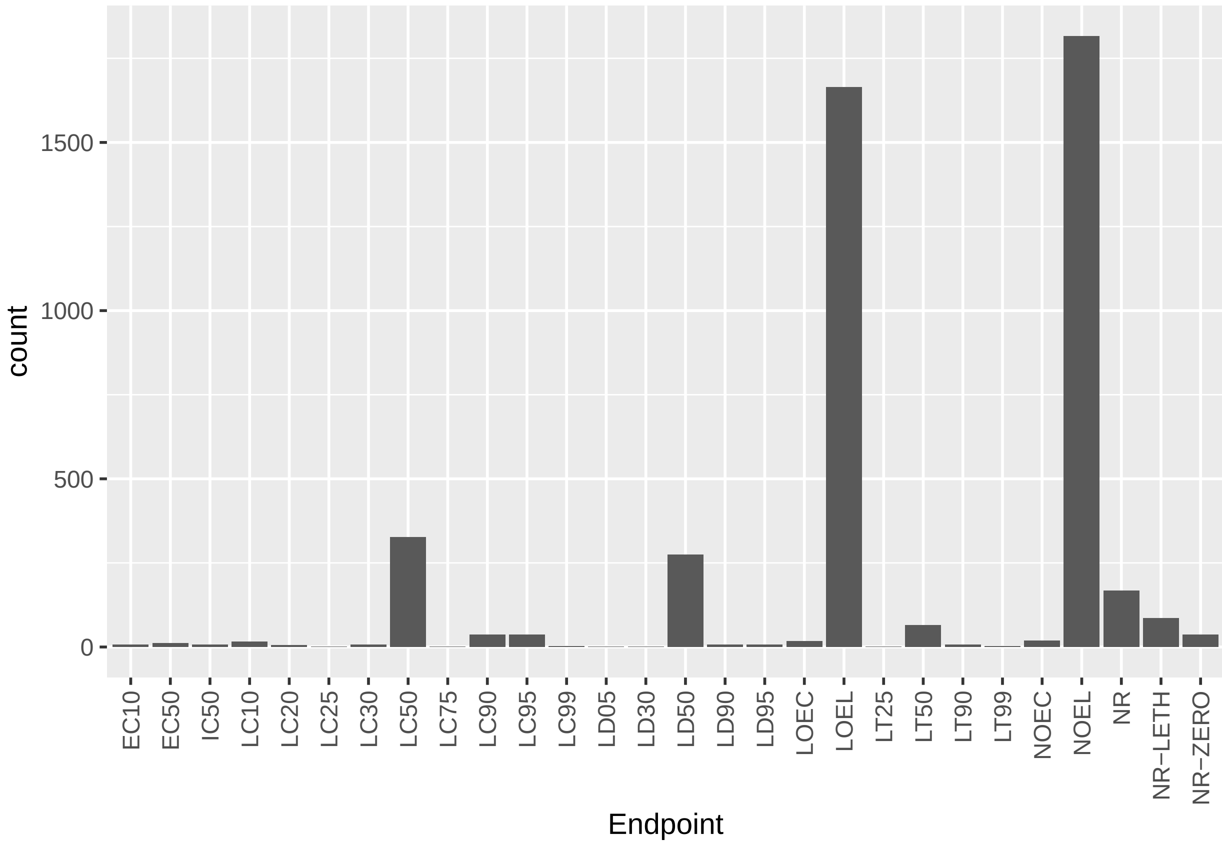

Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The lab and natural field locations tend to be the most common at different points in time. Most recetntly the lab has been the most common test location (since around 2010).

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




Answer: NOEL (terrestrial, No-observable-effect-level) and LOEL (terrestrial, Lowest-observable-effect-level) are the most common end points.

## 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) #class of column
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
#convert the data within this column to class of dates

class(Litter$collectDate) #checking the class of the new variable
```

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

```
unique(Litter$collectDat) #getting the unique values within the vector
```

```
## [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) #looking at unique results
```

```
##  [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
```

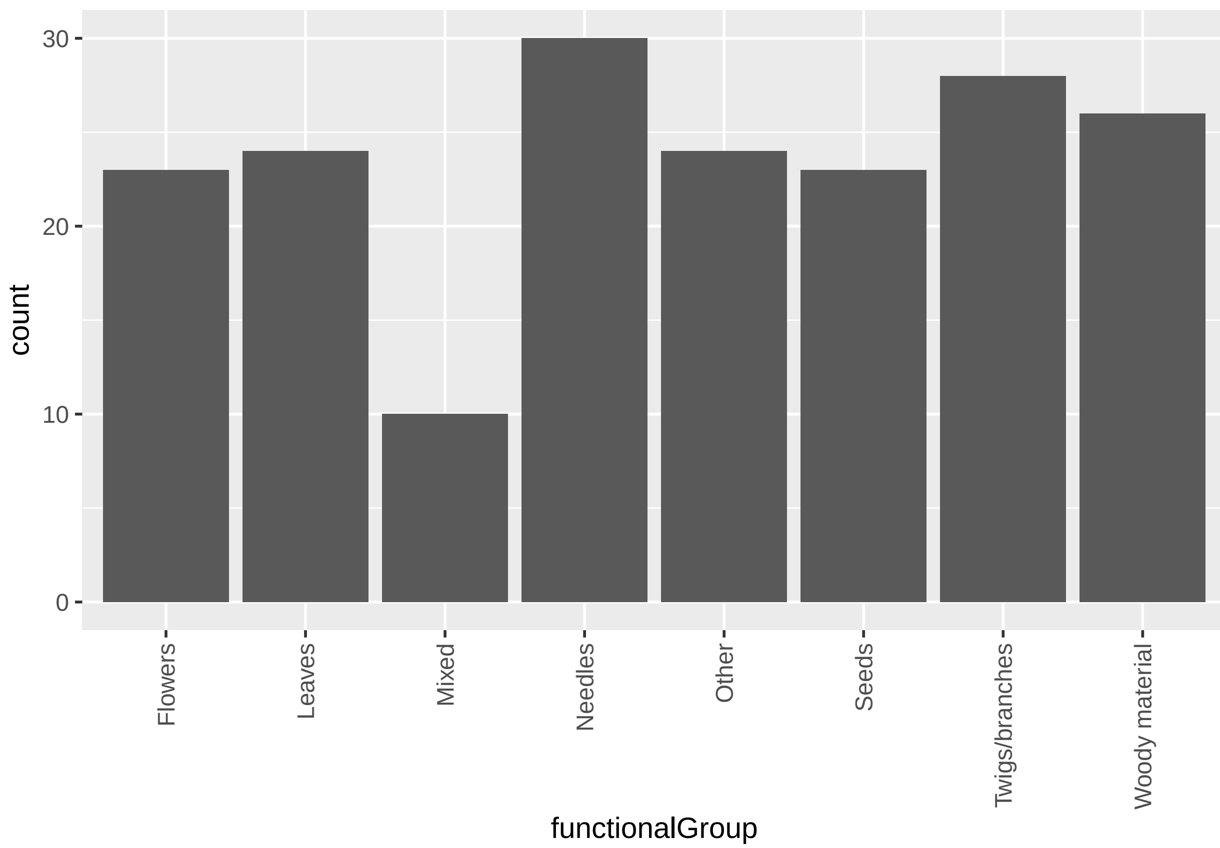
```
summary(Litter$plotID) #comparing to summary results
```

```
## 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 were sampled at Niwot Ridge. The unique function tells you all of the unique values within the column, whereas the summary function tells you the count of occurances of each unique value.

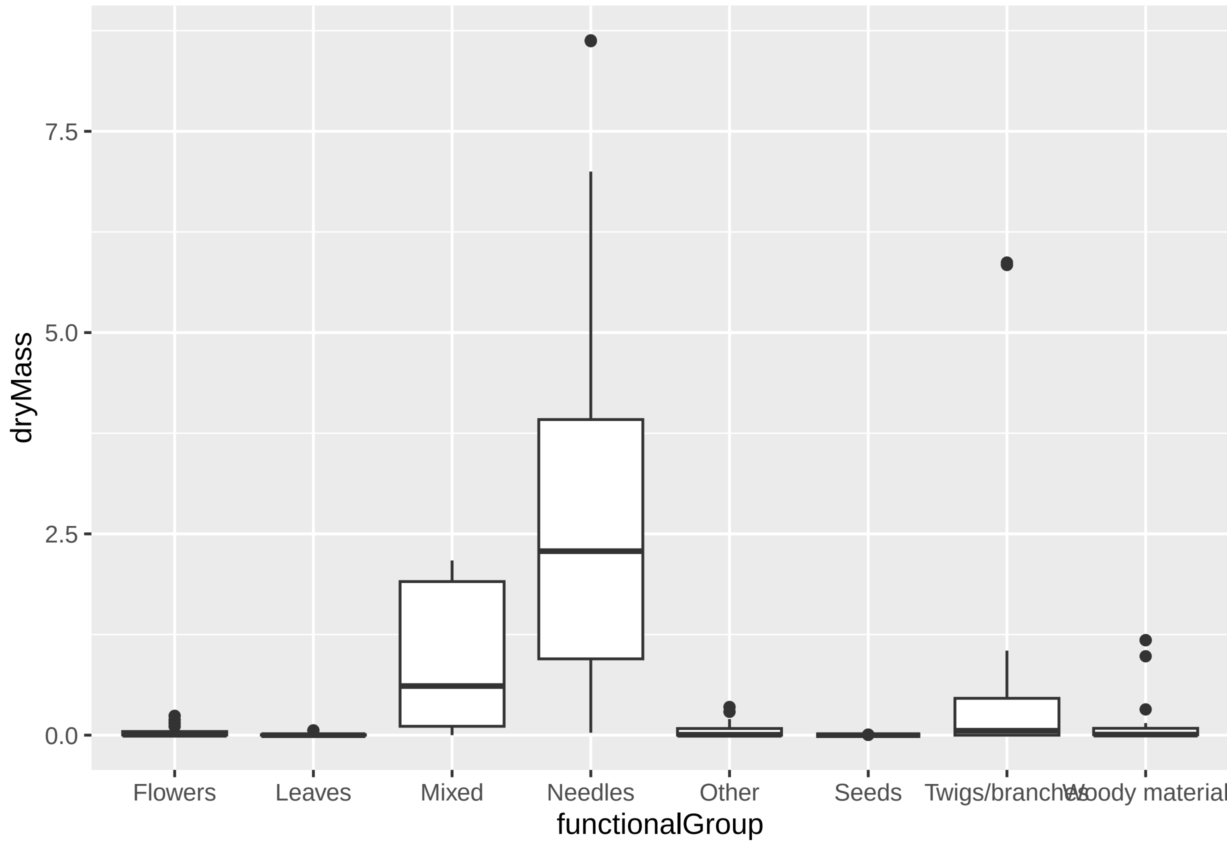
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

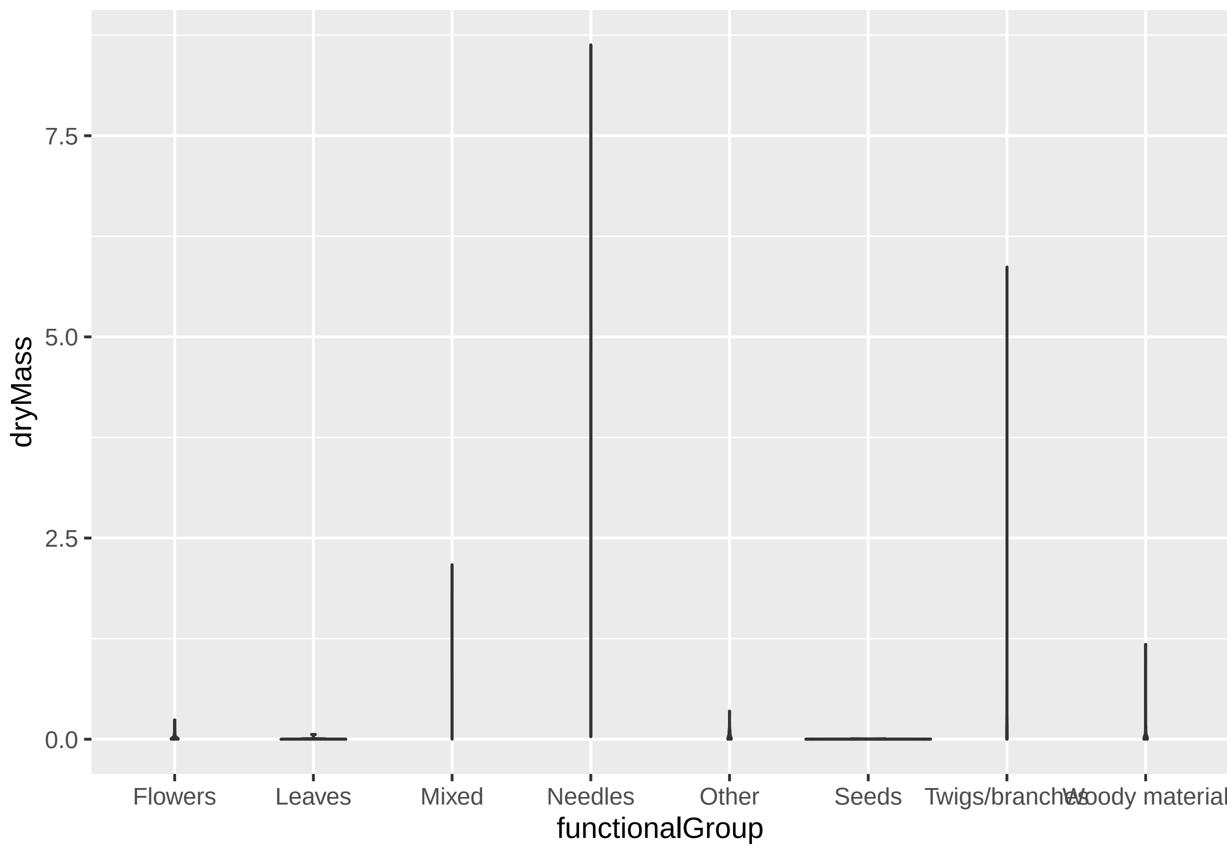



15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#box plot
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass, group = cut_width(functionalGroup, 1)))
```




```r
#violin plot
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 displays density distributions, but for the data we are working in there isn't a huge range of values for each functional group. As a result the violin plot isn't very informative and difficult to interpret. The boxplot is more interpretable; it shows us discernable values for dryMass for the different functional groups, most notably for the groups that had larger masses (mixed and needles).

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

> Answer: Needles tend to have the highest biomass at these sites.