# Assignment 3: Data Exploration

Jonathan Joyner

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.

```
#load lubridate and tidyverse packages using packets tab in lower right of Rstudio screen
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
#install.packages("tidyverse")
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 purrr   1.0.2     v tidyr   1.3.0

## -- 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
```

```r
#import and name Litter + Neonics datasets from Data/Raw folder with strings read as factors
Litter=read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv",stringsAsFactors = TRUE)
Neonics=read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_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: While neonicotinoids protect crops from predation by insects, they kill many insects that contribute to pollination and sustenance of other animals and plants in the environment. Examining data on neonicotonoids is key to understanding the external economic and social cost of their use.

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: It is important to understand the microbiology of the soil and how debris from surrounding plants and trees contributes to the ecosystem. In the case of Niwot Ridge, changes in climate at high altitude ecosystems can allow or prevent different forms of life from flourishing and create vegetative carbon fluxes over time.

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.) Debris (of up to 2x50 centimeters) is collected from both elevated and ground trabps to an accuracy of 0.01 grams (2.) Elevated traps are 0.5 meters square baskets and 80 centimeters off the ground. Ground traps are 3 meter by 0.5 meter triangular areas. (3.) There are 1-4 traps placed across 30 plots with vegetation greater than 2 meters. 26 of these plots are 26 by 26 meters and 4 are 40 by 40 meters.

## Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#Use dim() to find dimensions of Neonics dataset
dim(Neonics)
```

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

```
#Neonics is 4623 rows x 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?

```
#Use summary() to view effects and determine most frequently referenced
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 top three most common effects in order are Population, Mortality, and Behavior

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...]

```
#Use summary() and sort() to find the most common name species
summary(sort(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 most commonly studied species in order are the Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. All of these species belong to the order Hymenoptera. They are important pollinators for many plants and they have a large impact on ecosystem health.

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?

```
#Use class() on Neonics dataset to establish class of "Conc.1..Author."
class(Neonics$Conc.1..Author.)
```

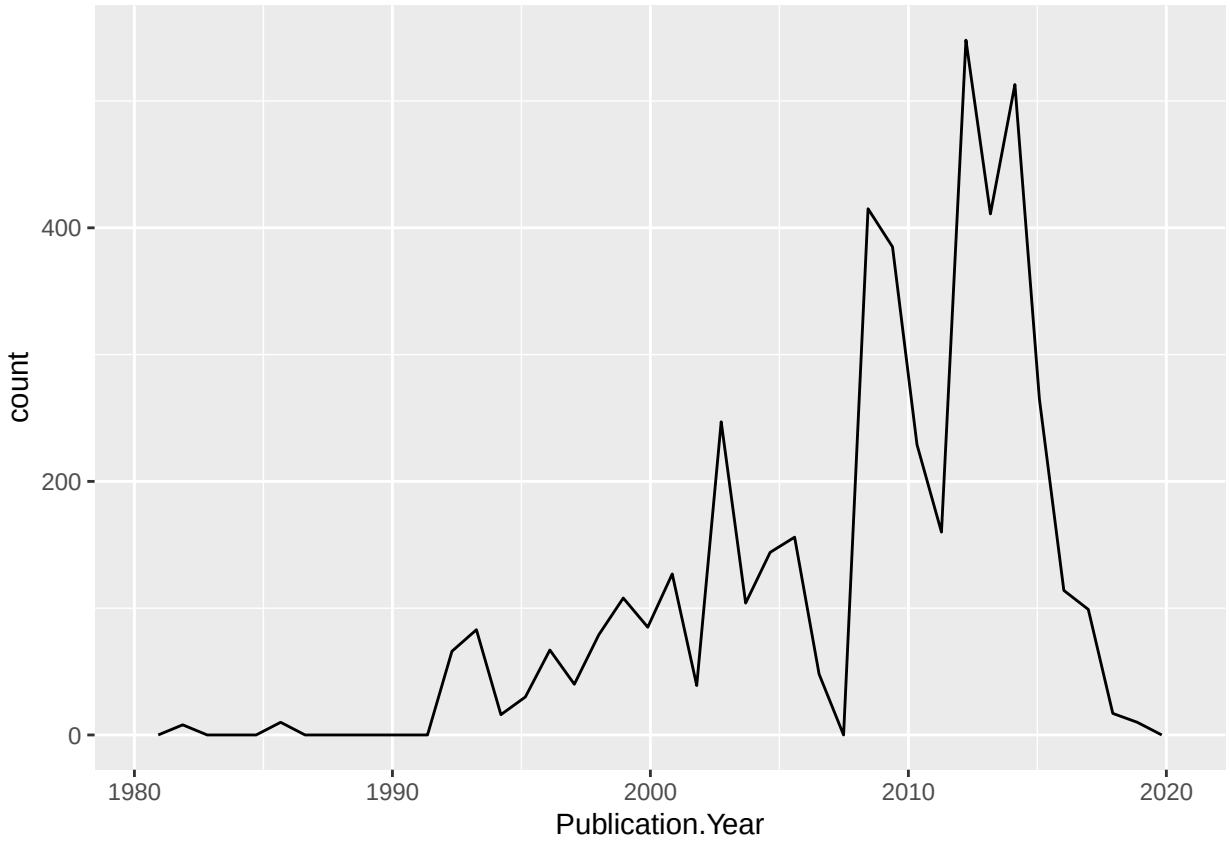
```
## [1] "factor"
```

Answer: The class of Conc.1..Author. is not numeric, but factor. This is the case because there is some data in the column which cannot be categorized as numeric.

## Explore your data graphically (Neonics)

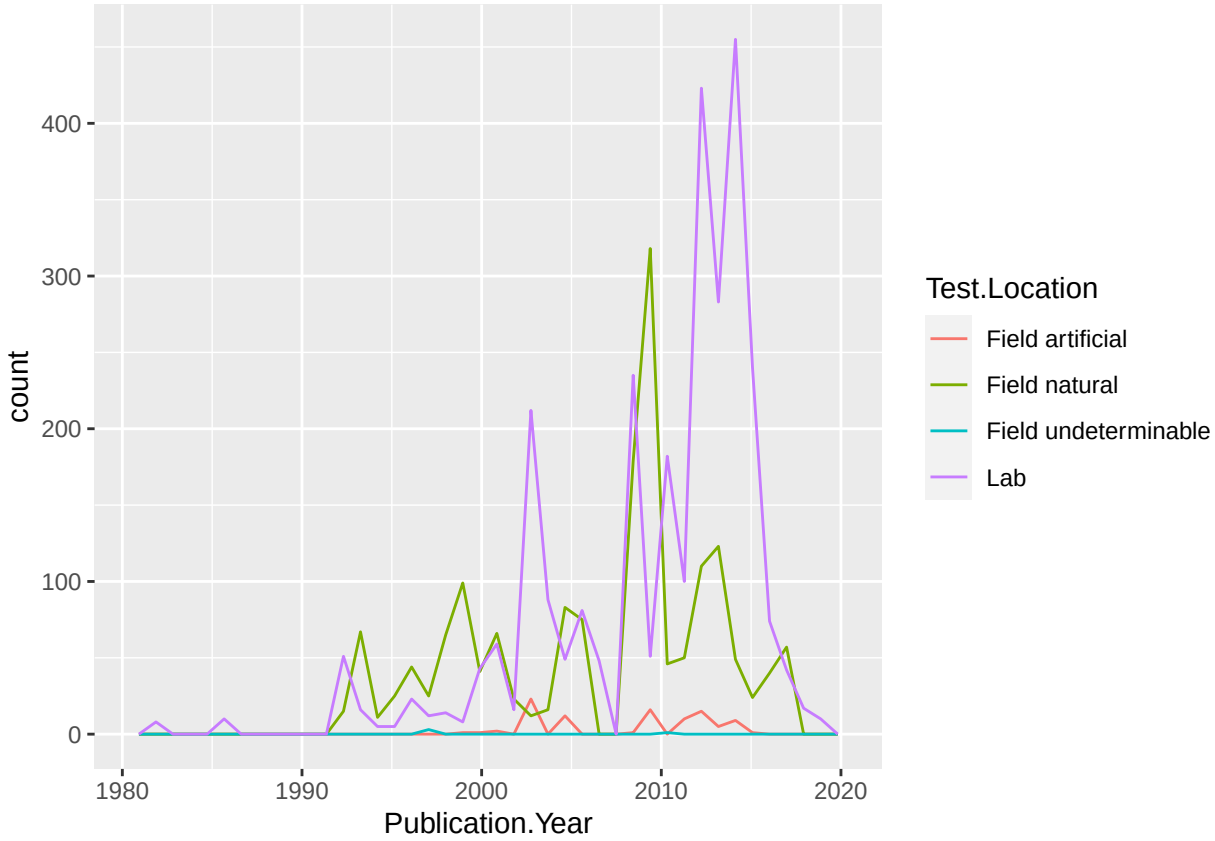
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
#Use ggplot() and geom_freqpoly to plot number of studies by publication year between 1980 and 2020
ggplot(Neonics) +
  geom_freqpoly(aes(Publication.Year), bins=40)
```




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

```
#Add color coded differentiation by test location
ggplot(Neonics) +
  geom_freqpoly(aes(Publication.Year,color=Test.Location), bins=40,)
```



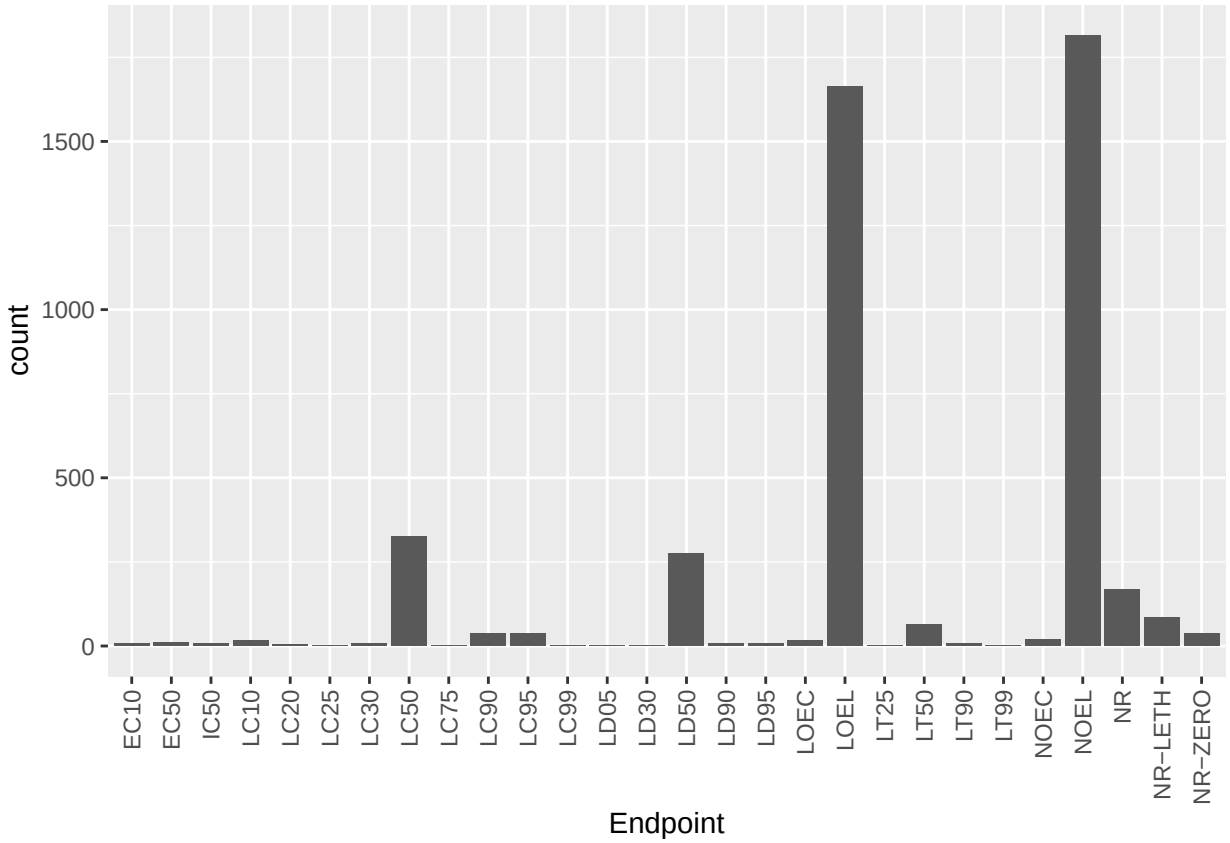

Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The two most common locations are the lab and the natural field. These two locations change places in popularity at certain intervals of time. For instance, the lab is most popular from around 2002-2008 and 2011-2017 whereas the natural field is most popular between 1993-1998 and during a short period from 2009-10.

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...]

```
#Create bar plot of Endpoints and use *tip* to rotate axis labels
ggplot(Neonics) +
  geom_bar(aes(Endpoint)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```




Answer: The two most common end points are NOEL (high concentration producing no observable effect level difference from controls) and LOEL (lowest observable effect level showing low concentrations producing significant difference from 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.

```
#Determine class of Litter data set column collectDate
class(Litter$collectDate)
```

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

```
#Change class from factor to date using as.Date
Litter$collectDate <- as.Date(Litter$collectDate)
#Confirm change in class of collectDate from factor to date
class(Litter$collectDate)
```

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

```
#Use unique function to find distinct values in collectDate column and determine dates litter was sampl
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`?

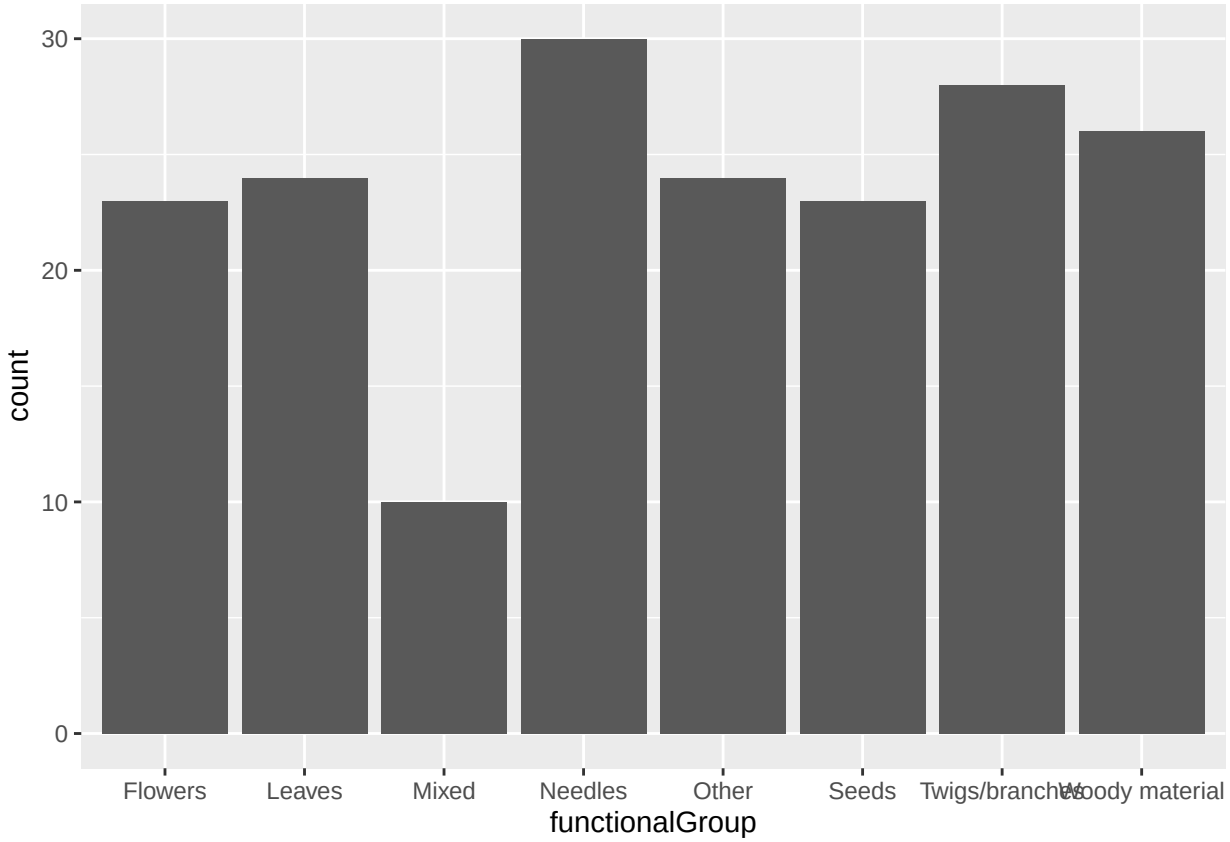
```
#Use unique function to determine unique values in plotID column and identify the range of plots used f
unique(Litter$plotID)
```

```
##  [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: Unique will determine the number of unique values including a total whereas summary will display the number of values along with the amount if times each value was identified.

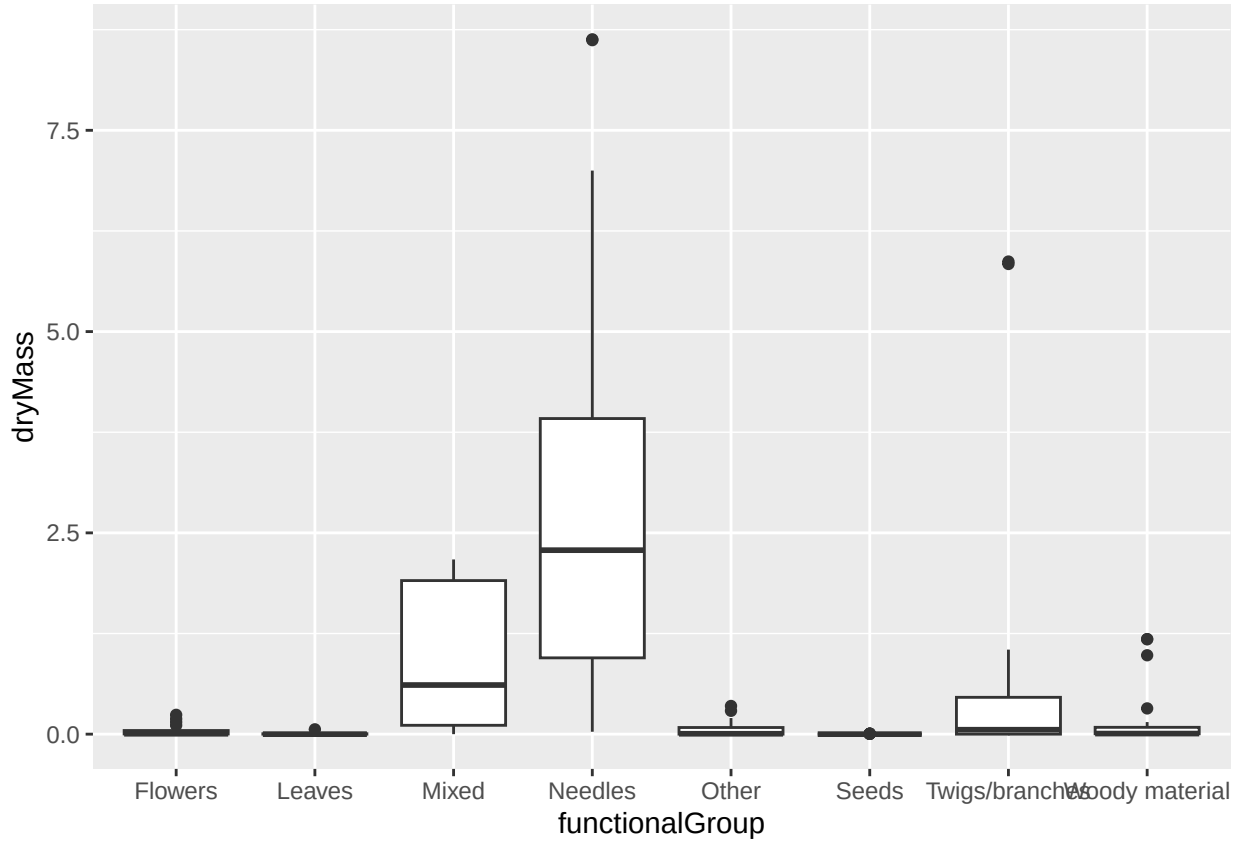
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#Create a bar graph which visualizes the distribution of samples by functional group
ggplot(Litter,aes(x=functionalGroup)) + geom_bar()
```

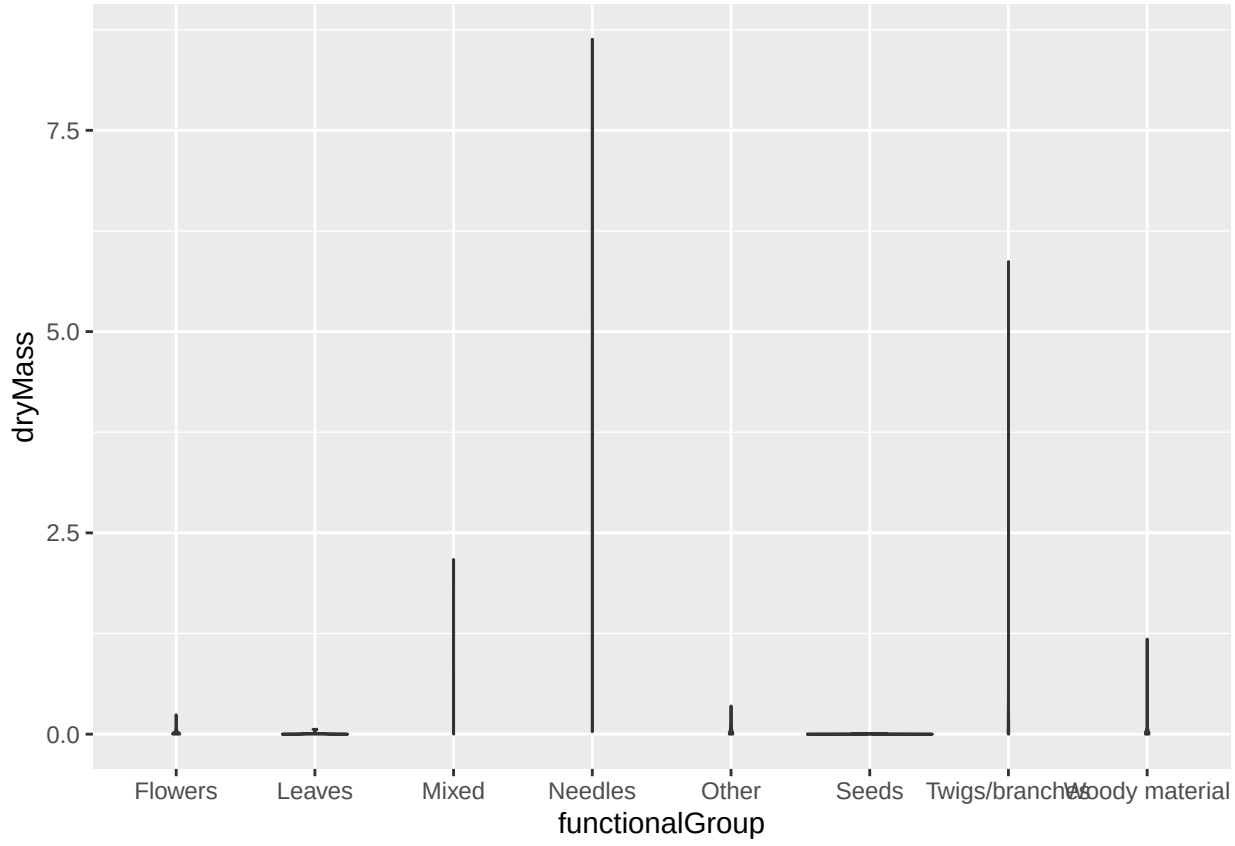



15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#Create a box and whisker plot with dryMass amounts on the y axis and group type on the y axis
ggplot(Litter) +
  geom_boxplot(aes(x=functionalGroup, y=dryMass))
```




```
#Create a violin plot with dryMass amounts on the y axis and group type on the y axis
ggplot(Litter) +
  geom_violin(aes(x=functionalGroup, y=dryMass))
```




```
draw_qartiles=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 boxplot shows differences in the mean and median along with outliers whereas the violin plot is not able to show density because most samples have minute differences in value.

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

> Answer: On average, needles have the highest biomass, but a small number of twigs also have high biomass.