# Assignment 3: Data Exploration

Blair Johnson, Section 01

## OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Exploration.

### Directions

1. Change "Student Name, Section #" on line 3 (above) with your name and section number.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "FirstLast_A03_DataExploration.Rmd") prior to submission.

The completed exercise is due on <>.

### 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 add the `stringsAsFactors = TRUE` parameter to the function when reading in the CSV files.**

```
library(tidyverse)
library(ggplot2)

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 ecotoxicologoy of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: This dataset examining the ecotoxicology of neonicotinoids on insects can help inform future ecotoxicology research by allowing the EPA to better understand the effects on insecticides on insects and how varying levels of exposure to neoincotinoids can affect the lifespan of an insect. This can then help scientists better understand the impacts of particular insecticides on insects and how particular types of neonicotinoids may be more or less potent.

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 may be interested in studying litter and woody debris to better understand their impact on ecosystem health as well as the cycling of carbon and other nutrients across the terrestrial sites.

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: Litter and woody debris are sampled at terrestrial NEON sites with woody vegetation in tower plots. The sampling occurs in a different number of plots depending on characteristics of the vegetation. Information on sampling methods *Trap placements within the plots are randomized when the sites contain over 50% aerial cover of woody vegetation over 2 meters in height.* Trap placements within plots are targeted when the sites contain less than 50% aerial cover of woody vegetation. *The sampling frequency of sites varies based on the type of site. Ground traps are sampled once per year whereas elevated traps are sampled more frequently (approximately once every two weeks).

## Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

```
#4623 rows by 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?

```
summary(Neonics$Effect) #Shows which effects are the most prevalent
```

```
##     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 believe the effects may be specifically of interest because these effects demonstrate how the neonics may affect the life expectancy of species being studied and how the neonics may place a role in insect mortality.

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 in the dataset are the following: 1) Honey Bee, 2) Parasitic Wasp 3) Buff tailed bumblebee 4) Carniolan Honey Bee 5)Bumble Bee 6)Italian Honeybee. With the exception of the parasitic wasp, all of these species are bees. However, all six of these species are pollinators which demonstrate that pollinators are of interest moreso than other species. Since pollinators are crucial in plant fertilization, the EPA may be more interested in studying the effects of neonics on these species in order to understand the chemicals potential impact on plant health.

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"
```

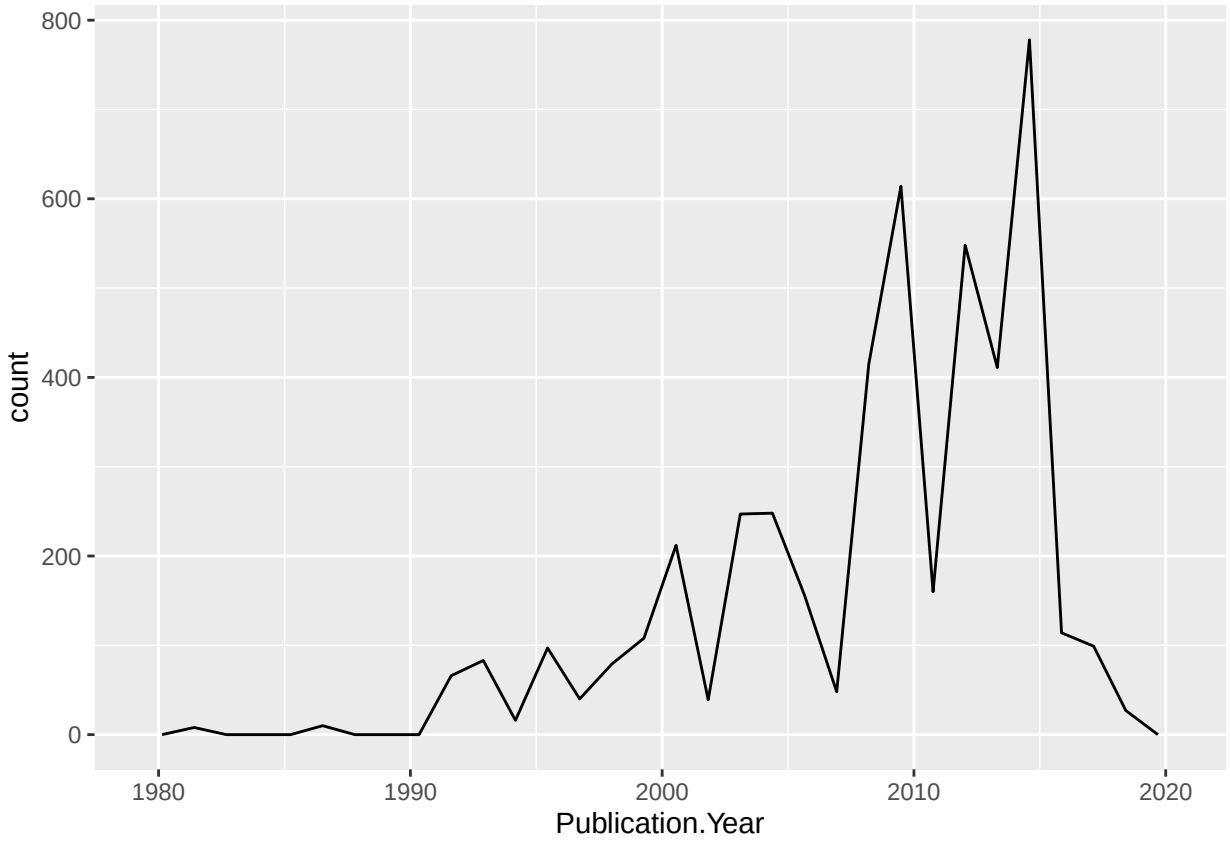
Answer: The class of Conc.1...Author is a factor. This is because some of the concentrations are not numeric (ex. NR which may be not reported) or are numeric values with a "~" sign which indicate that they are approximations.

## Explore your data graphically (Neonics)

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

```
library(ggplot2)
ggplot(Neonics) + geom_freqpoly(aes(x=Publication.Year))
```

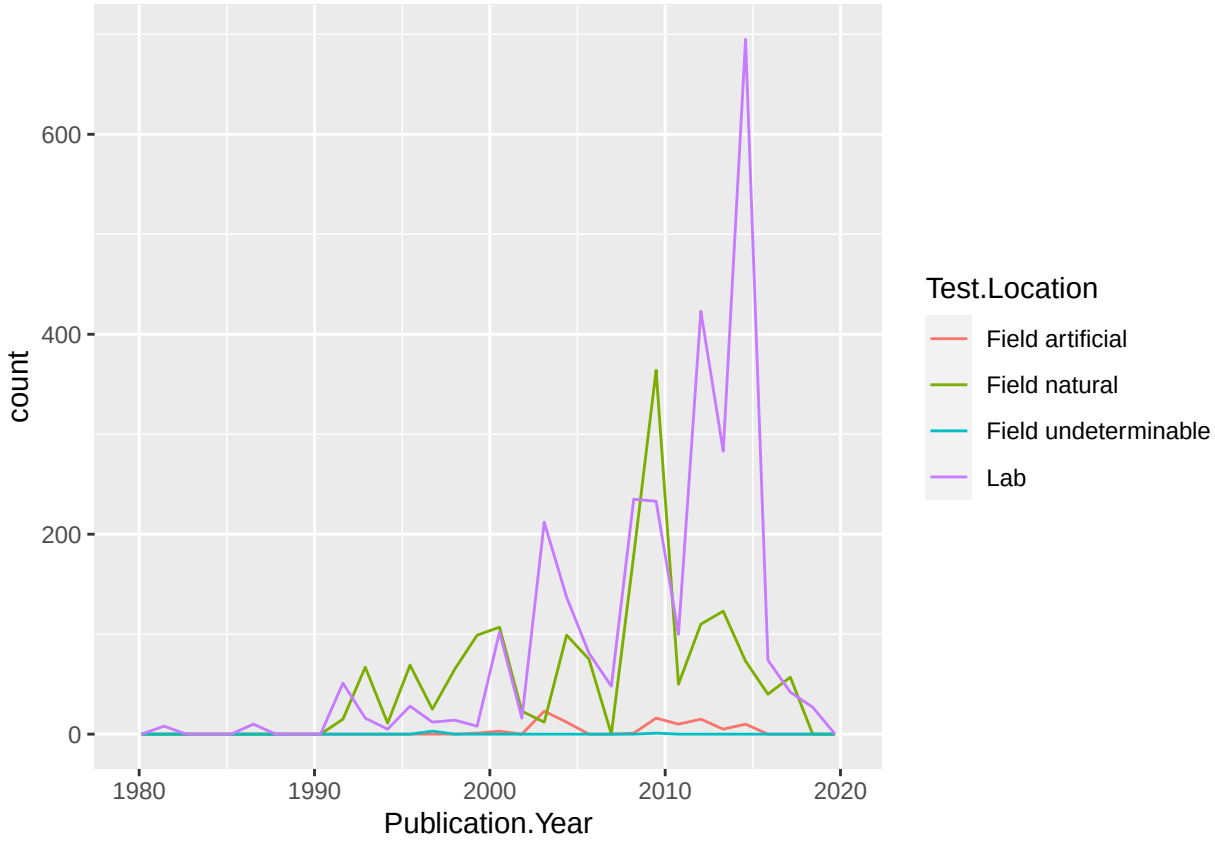
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




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

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

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

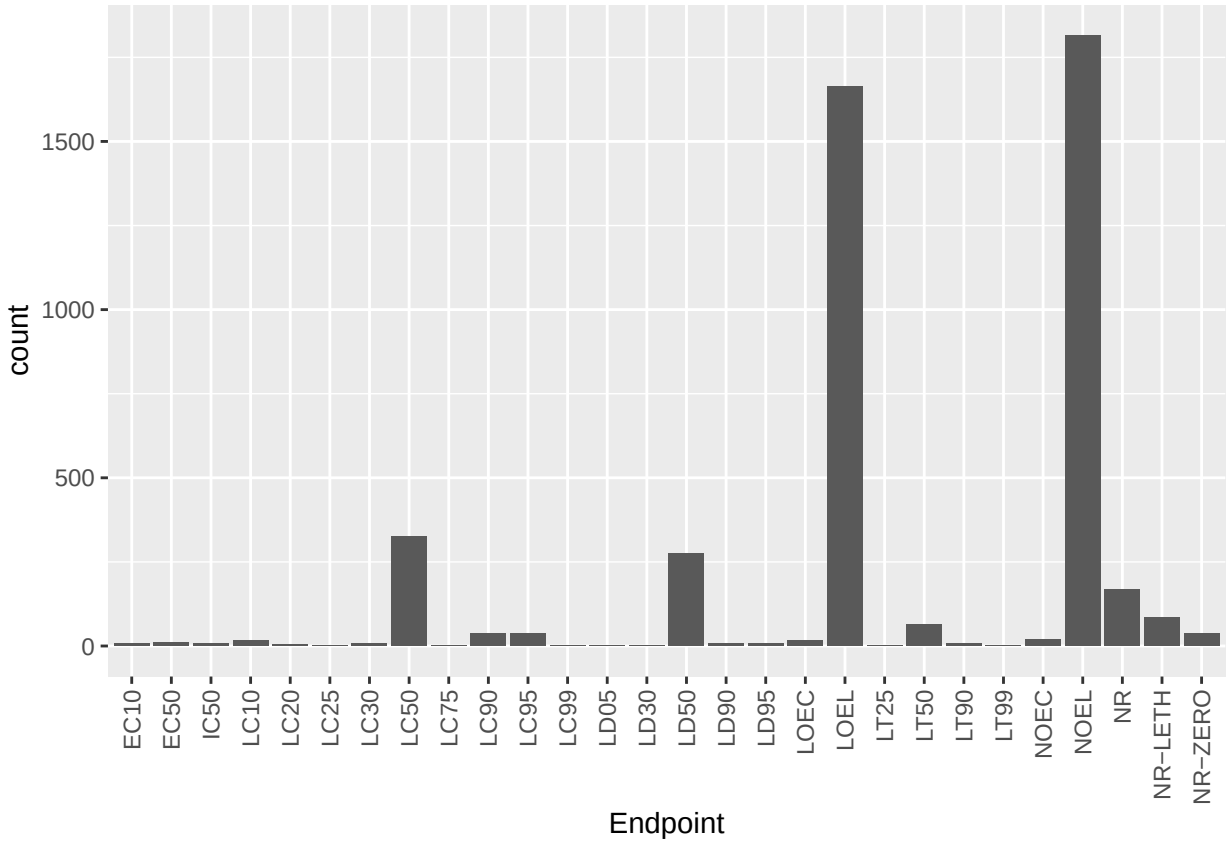



Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: By examining the graph, it is evident that the most common test locations from 1990 to present day are a lab and a field-natural. The field-natural is the most prevalent test location from the early 1990s to early 2000s while the lab surpasses the field-natural in frequency starting in the early 2000s. Based on this graph, the labs' usage increases rapidly and peaks between 2010-2015 and begins to sharply decline starting in 2015. Prior to the peak in lab testing locations, the field-natural experiences a peak followed by a decrease.

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.

```
ggplot((Neonics), na.rm=TRUE) +geom_bar(aes(x=Endpoint)) +theme(axis.text.x = element_text(angle=90, vj
```




Answer: The most common endpoint coordinates are NOEL and LOEL. NOEL corresponds to no-observable-effect-level and LOEL corresponds to the lowest-observable-effect-level. Therefore, the more commonly observed endpoints in this dataset correspond to the highest dose where there is no effect on the control group and the lowest dose where there is an effect on the group control. These endpoints can help one better understand the threshold at which a neonics has an effect on insects.

## 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) #collectDate is a factor
```

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

```
library(lubridate)
```

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

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

```
Litter$collectDate<-as.Date(Litter$collectDate) #collectDate is now a date

unique(Litter$collectDate)
```

```
## [1] "2018-08-02" "2018-08-30"
```

```
#Litter was sampled on August 2 and August 30 in 2018
```

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) #12 plots were sampled at Niwot Ridge
```

```
##  [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
```

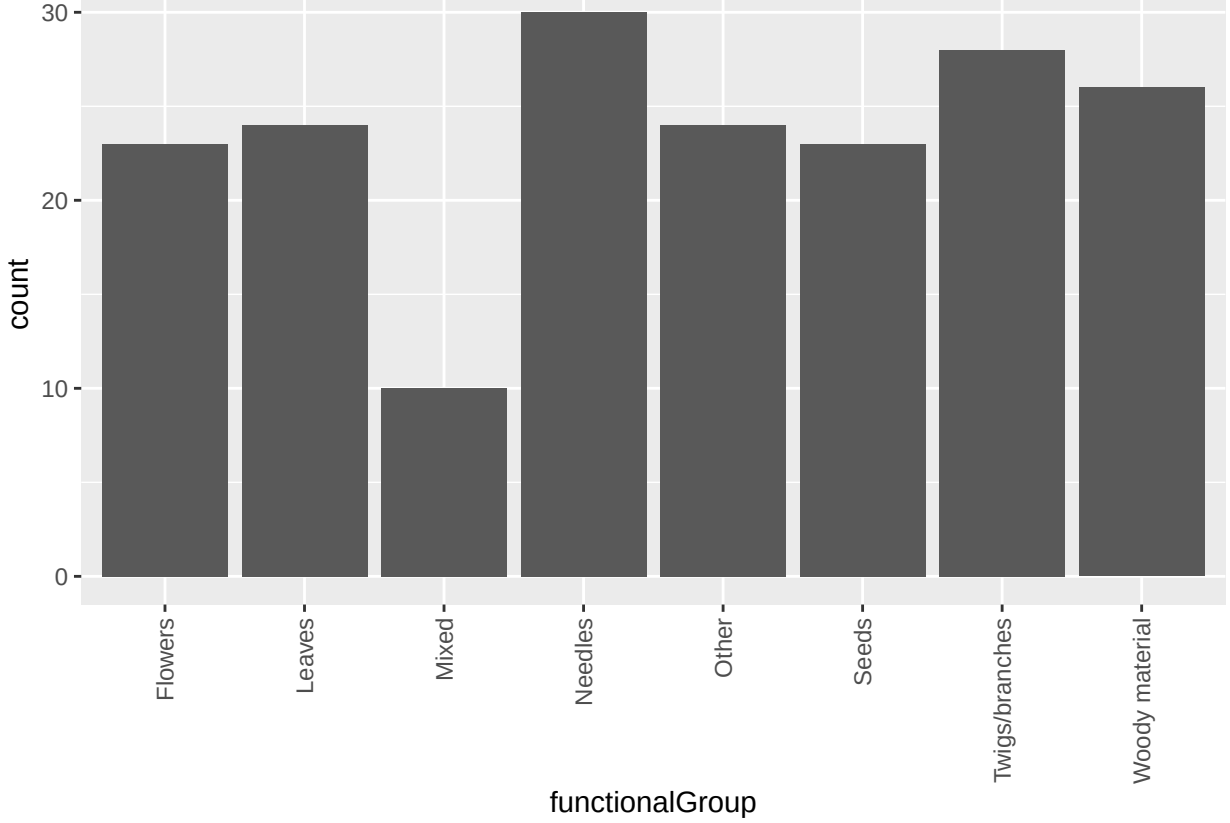
```
summary(Litter$plotID)
```

```
## 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: By using the "unique" function, we can determine that 12 plots were sampled at Niwot Ridge. When obtaining data on the plots using the summary function, we obtain the number of observations at each unique plot rather than the number of unique plots being sampled in the dataset.

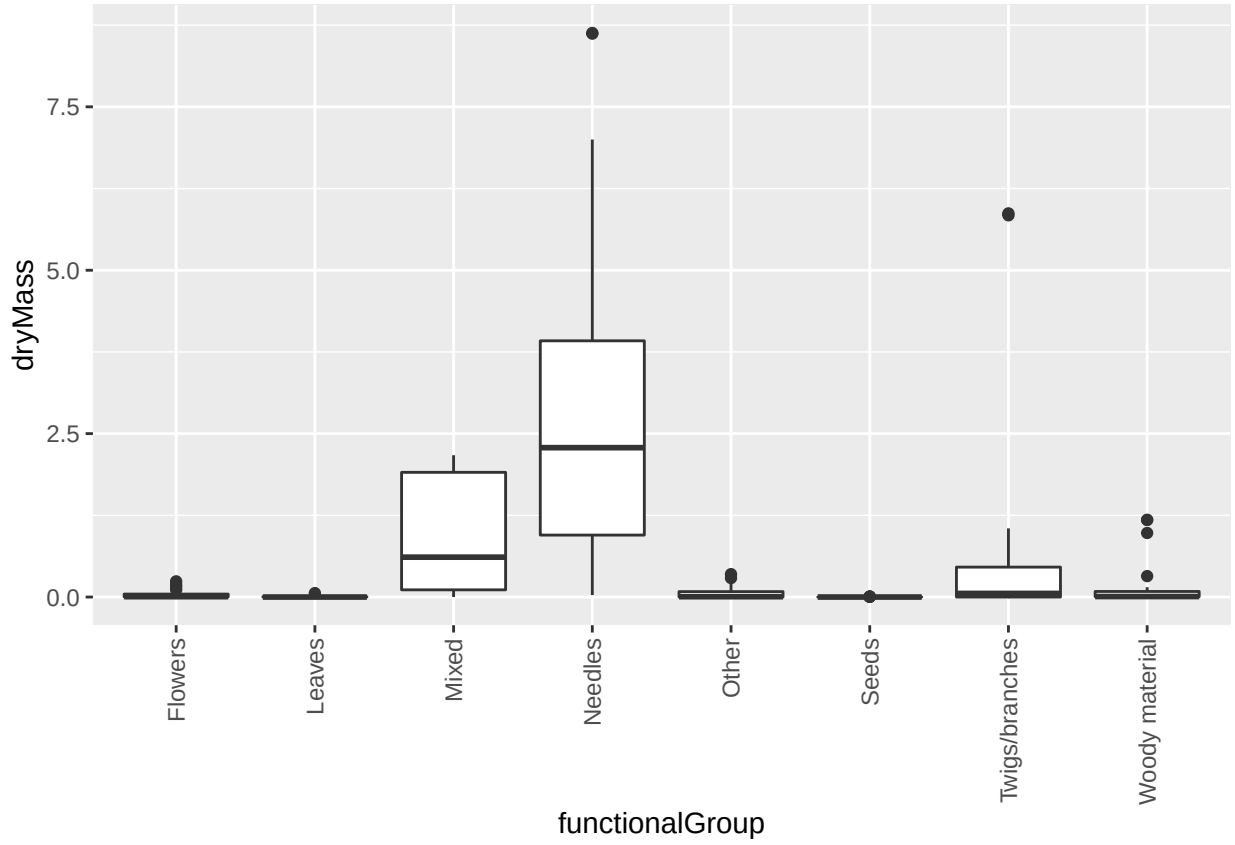
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter) + geom_bar(aes(x=functionalGroup)) + theme(axis.text.x = element_text(angle=90, vjust=0.5
```

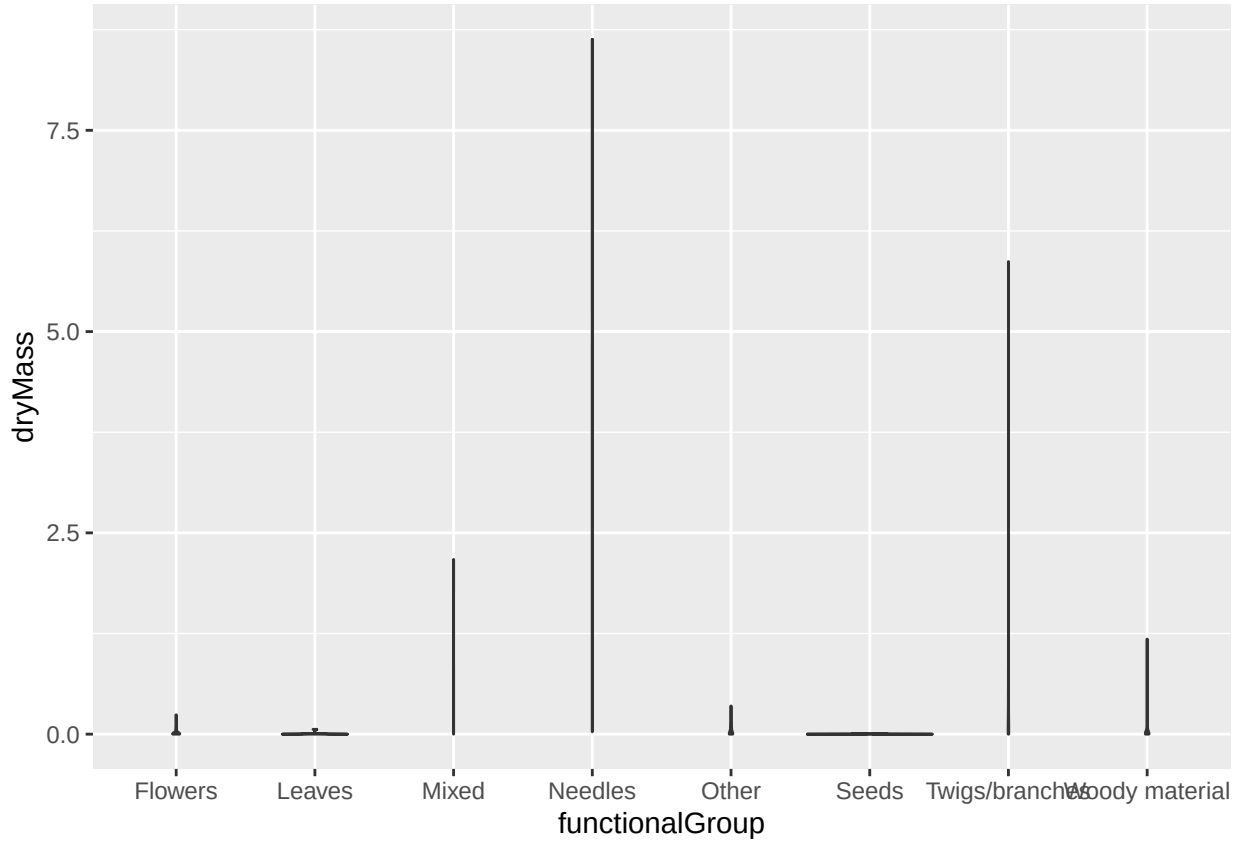


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(Litter) +
  geom_boxplot(aes(x=functionalGroup, y= dryMass)) + theme(axis.text.x = element_text(angle=90, vjust=0
```




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




```
#Tried to visualize data with the violin plot. The violin does not show spread as easily
```

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

> Answer: In this case, the boxplot is a more effective visualization option than the violin plot because it displays the spread of data and provides a visualization of the dryMass summary statistics. The boxplot displays the median, interquartile range, and outliers across each functional group and those statistics are not apparent in the violin plot.

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 at the sites in this study. While twigs and branches tend to have a lower biomass relative to needles and mixed litter, there is an outlier that is heavier.