

# Assignment 3: Data Exploration

Jordan Mullens

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.

```
#Checking working directory
getwd()
```

```
## [1] "/home/guest/EDA-Spring2023"
```

```
setwd("/home/guest/EDA-Spring2023")
```

```
#loading necessary packages
library(tidyverse)
library(lubridate)
```

```
#uploading two data sets
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: Humans perilously overuse pesticides, and it's possible that these pesticides have non-target effects. Researchers have found that neonicotinoids can move from treated plants to pollinators and other insects. Herbivores and omnivorous species die after directly consuming plant tissues containing lethal neonicotinoid concentrations. Nonsusceptible herbivores encounter neonicotinoids as they feed. As these individuals persist in the food web, they expose other consumers to the insecticide. These chemicals pose broader risks to biodiversity and food webs than previously recognized.

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: Plant debris is a component of carbon storage in forest ecosystems and recycling of nutrients. Forest age directly determines storage accumulation of plant debris and insights into the dynamics of plant debris may help forest managers understand the impact of management regimes on material cycles and energy flow of forest ecosystems. Additionally, litter and woody debris may be home to the insects/ invertebrates exposed to neonicotinoids.

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. Locations of tower plots are selected randomly within the 90% flux footprint of the primary and secondary airsheds (and additional areas in close proximity to the airshed, as necessary to accommodate sufficient spacing between plots). In sites with forested tower airsheds, the litter sampling is targeted to take place in 20 40m x 40m plots. In sites with low-statured vegetation over the tower airsheds, litter sampling is targeted to take place in 4 40m x 40m tower plots (to accommodate co-located soil sampling) plus 26 20m x 20m plots. 2. One litter trap pair (one elevated trap and one ground trap) is deployed for every 400 m<sup>2</sup> plot area, resulting in 1-4 trap pairs per plot. Trap placement within plots may be either targeted or randomized, depending on the vegetation. In sites with > 50% aerial cover of woody vegetation > 2m in height, placement of litter traps is random and utilizes the randomized list of grid cell locations being utilized for herbaceous clip harvest and bryophyte sampling. 3. Ground traps are sampled once per year. Target sampling frequency for elevated traps varies by vegetation present at the site, with frequent sampling (1x every 2 weeks) in deciduous forest sites during senescence, and infrequent year-round sampling (1x every 1-2 months) at evergreen sites

## Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

```
dim(Litter)
```

```
## [1] 188 19
```

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: This data may enable researchers to determine the most vulnerable species in the food web and understand how neonicotinoids may potentially impact humans through bioaccumulation.

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...]

```
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 Woolly 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: They are all pollinators, and their population decline could result in diminished agricultural production.

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(Neonics$Conc.1..Author.)
```

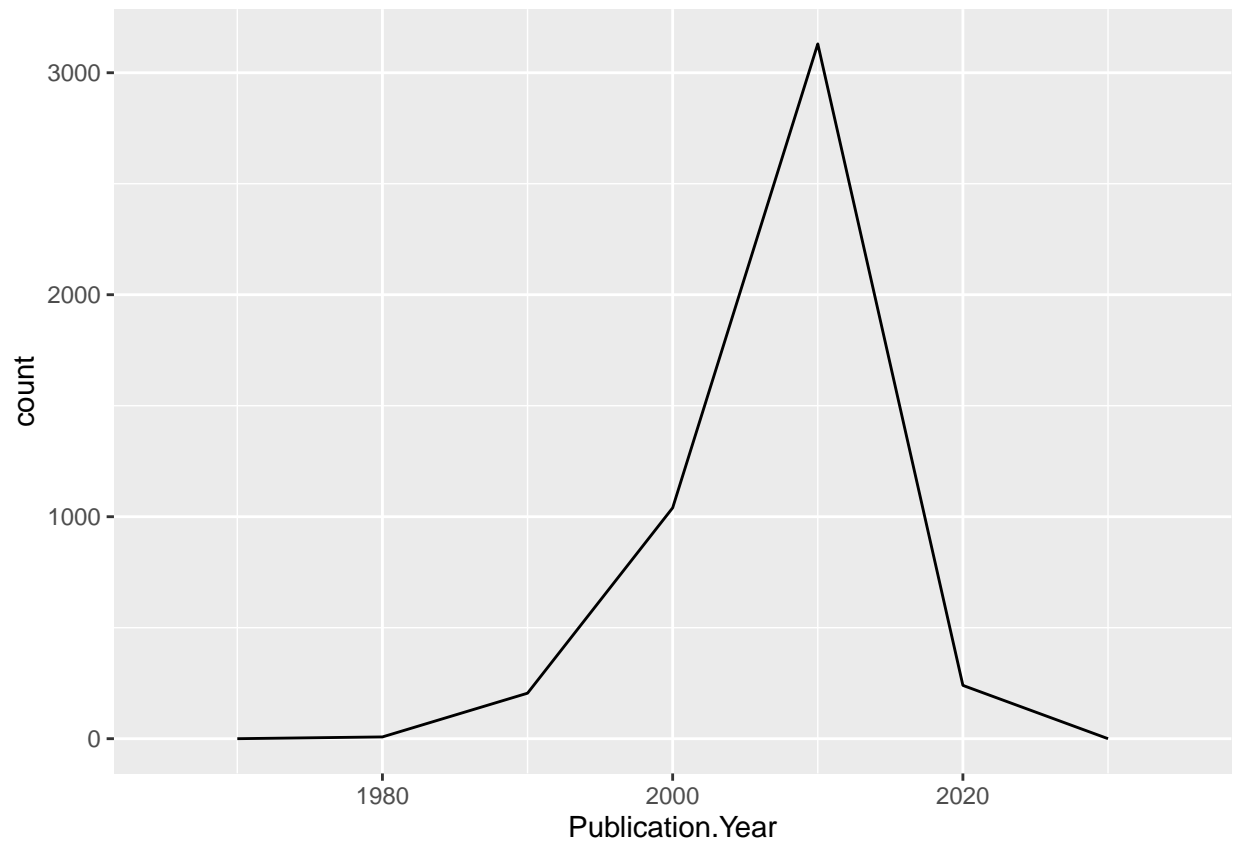
```
## [1] "factor"
```

Answer: It is a factor variable. It is not numeric because it is a variable that could be both a number or word.

## Explore your data graphically (Neonics)

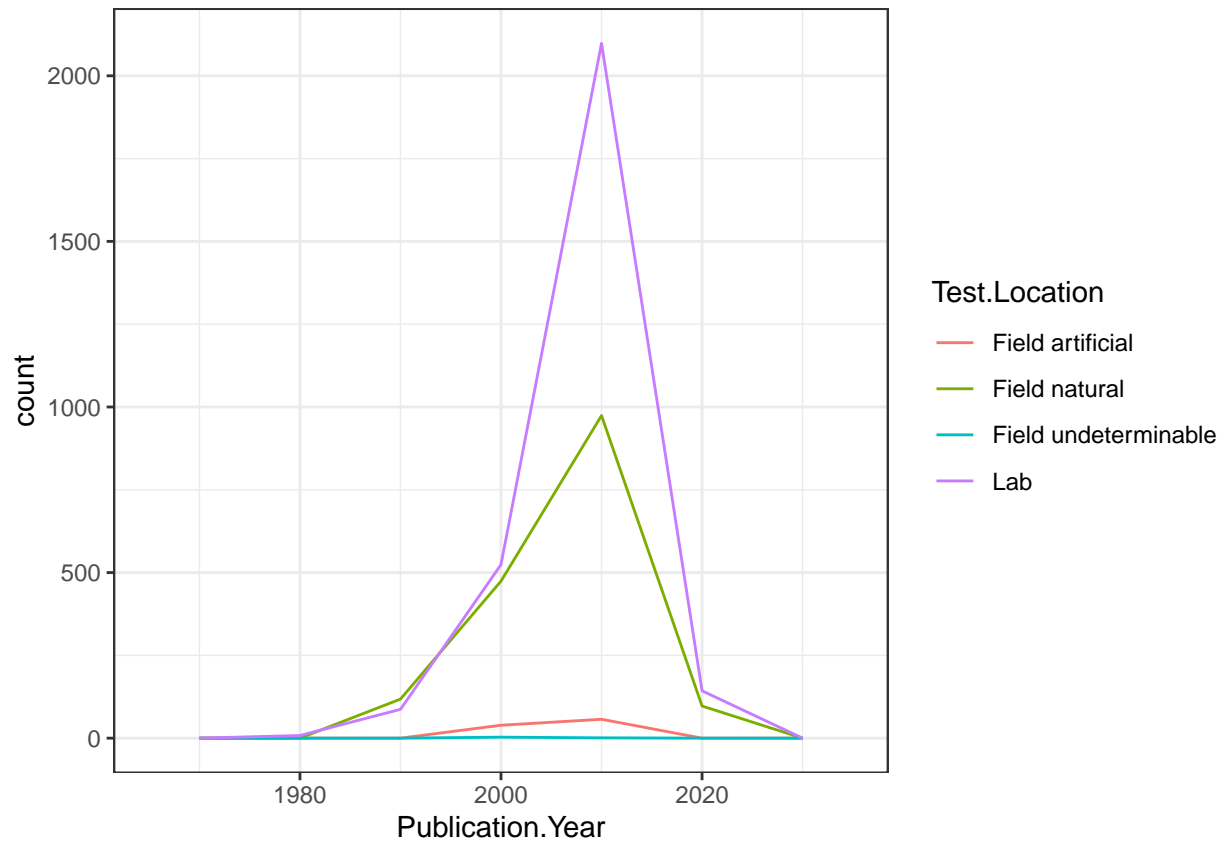
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
library(ggplot2)
ggplot(Neonics, aes(x=Publication.Year)) +
  geom_freqpoly(binwidth=10)
```



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

```
ggplot(Neonics, aes(x=Publication.Year, color=Test.Location)) +
  geom_freqpoly(binwidth=10) +
  theme_bw()
```



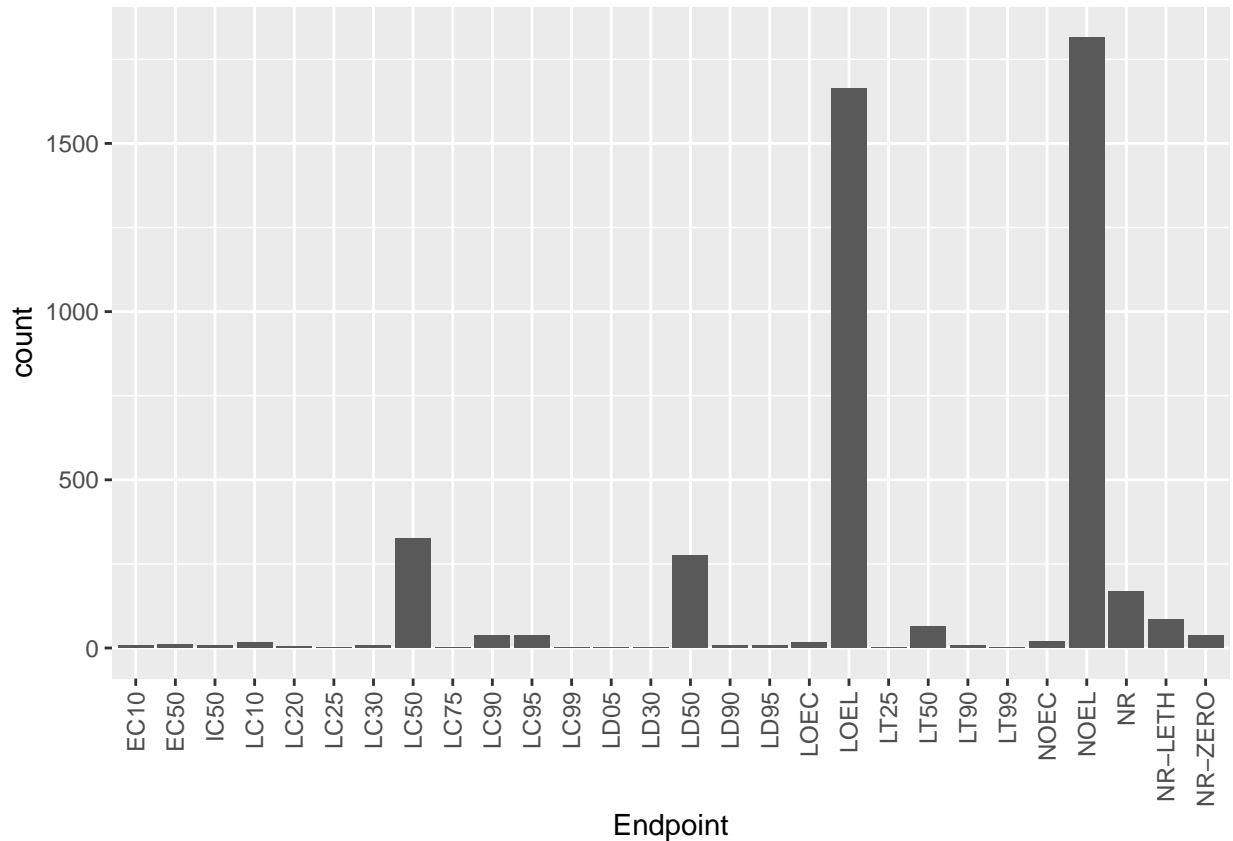
Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: The most common test locations are lab and field natural. The number of both increase over time.

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



Answer: The most common are LOEL and NOEL. LOEL=Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC). NOEL=No-observable-effect-level: highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test (NOEAL/NOEC).

## Explore your data (Litter)

- 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)
```

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

```
#change collectDate to a date instead of factor
CD.Collection.Date <- ymd(Litter$collectDate)
```

```
#check to ensure collectDate is now a date variable
class(CD.Collection.Date)
```

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



```
#Determine which dates litter was sampled in Aug 2018
unique(CD.Collection.Date)
```

```
## [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`?

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

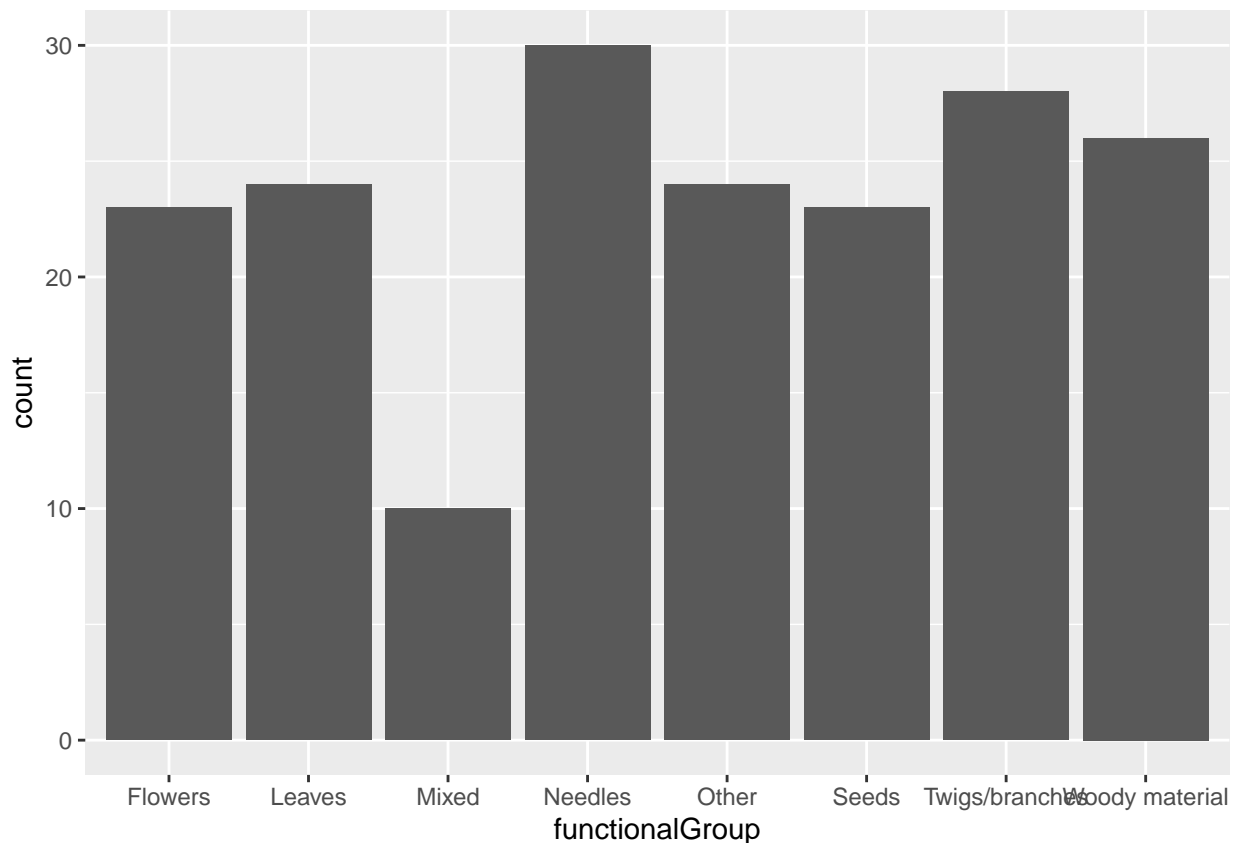
```
unique(Litter$plotID)
```

```
## [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: using the `unique` function removes duplicates in the data

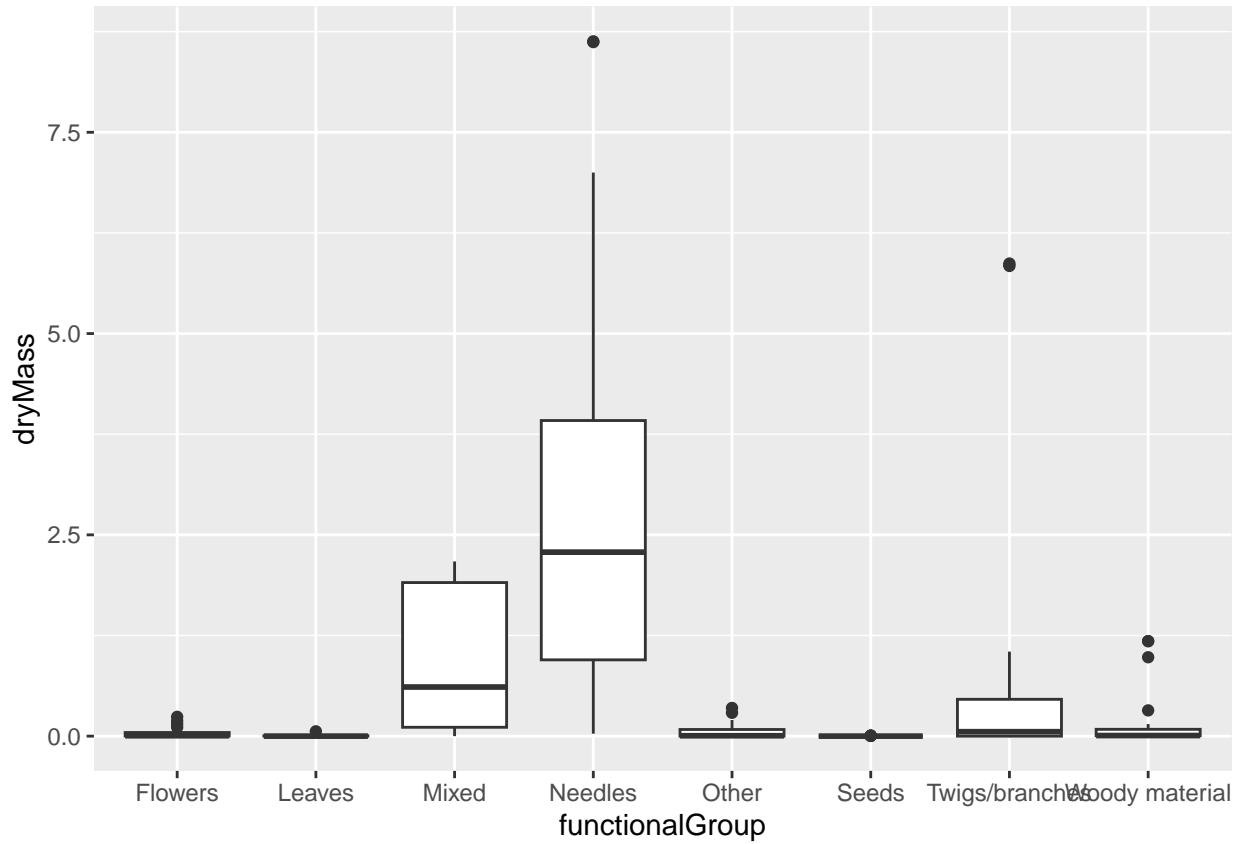
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

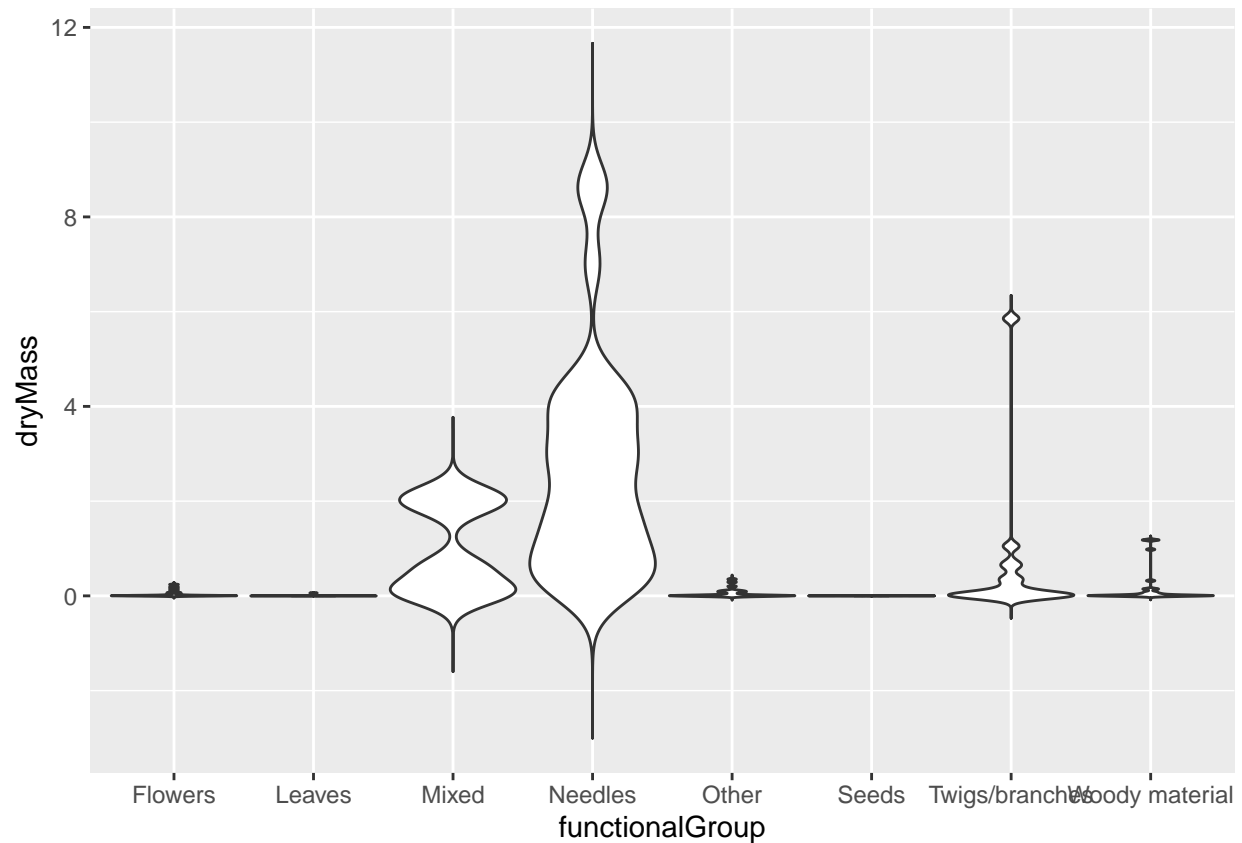


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of `dryMass` by functional-Group.

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



```
#set up violin plot with a maximum width to 1 for all violins, disabled trimming violins to the range of the data  
ggplot(Litter, aes(x=functionalGroup, y=dryMass)) +  
  geom_violin(scale = "width", trim = FALSE, adjust = 0.5)
```



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

Answer: Box plot is more valuable because we can see how the data is distributed and we can quickly detect outliers in the data. Violin plots are more useful when looking at the density of the data. Additionally, violin plots are more useful when the data is multimodal.

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

Answer: Needles