

# Assignment 3: Data Exploration

Peter Wasswa

Fall 2024

## 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 Canvas.

**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. Load necessary packages (tidyverse, lubridate, here), check your current working directory 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 sub-command to read strings in as factors.

```
#Set up: Check working directory  
getwd()
```

```
## [1] "/home/guest/Assignments/EDA_Spring2025_TA"
```

```
#Set up: Load libraries  
library(tidyverse)  
#Read datasets
```

```
Neonics <- read.csv('./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv', stringsAsFactors = TRUE)  
Litter <- read.csv('./Data/Raw/NIWO_Litter/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: Neonicotinoids persist in the environment for long durations. They can harm non-target species and, therefore, can cause unintended ecological effects. We might be interested in their ecotoxicology on insects in order to determine the insects that could be vulnerable to neonicotinoids.

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 forest litter and woody debris to study the relationship between climatic variables such as temperature, humidity etc and forest ecosystems. Decomposition of such litter is also crucial for nutrient cycles and, therefore, the amount of litter and woody debris might be of interest to foresters, ecologists and/or biogeochemists

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. A single observation of litter data is recorded from a single collection event and a single trap (1 spatial-temporal observation). 2. The number and placement (targeted or randomized) of elevated and ground traps deployed depends on physical features of vegetation 3. Frequency of sampling also depends on types of vegetation ranging from once every two-weeks to once a year.

## Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#Show dimensions of the Neonics 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? [Tip: The `sort()` command is useful for listing the values in order of magnitude...]

```
#Show summary of Effect Neonics dataset  
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: Population has the most records (1803), followed by Mortality (1493), and then Behavior at a distant third. This indicates that most studies just counted the number of individual spider/insects exposed, while the second one recorded only deaths, and the third looks more carefully for impacts on behavior

- 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: Explore the help on the `summary()` function, in particular the `maxsum` argument...]

```
#Show summary of Common name
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: Bees appear to be the most studied group, which makes sense given their importance to agriculture.

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? [Tip: Viewing the dataframe may be helpful...]

```
#Show class of `Conc.1..Author.`
class(Neonics$Conc.1..Author.)
```

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

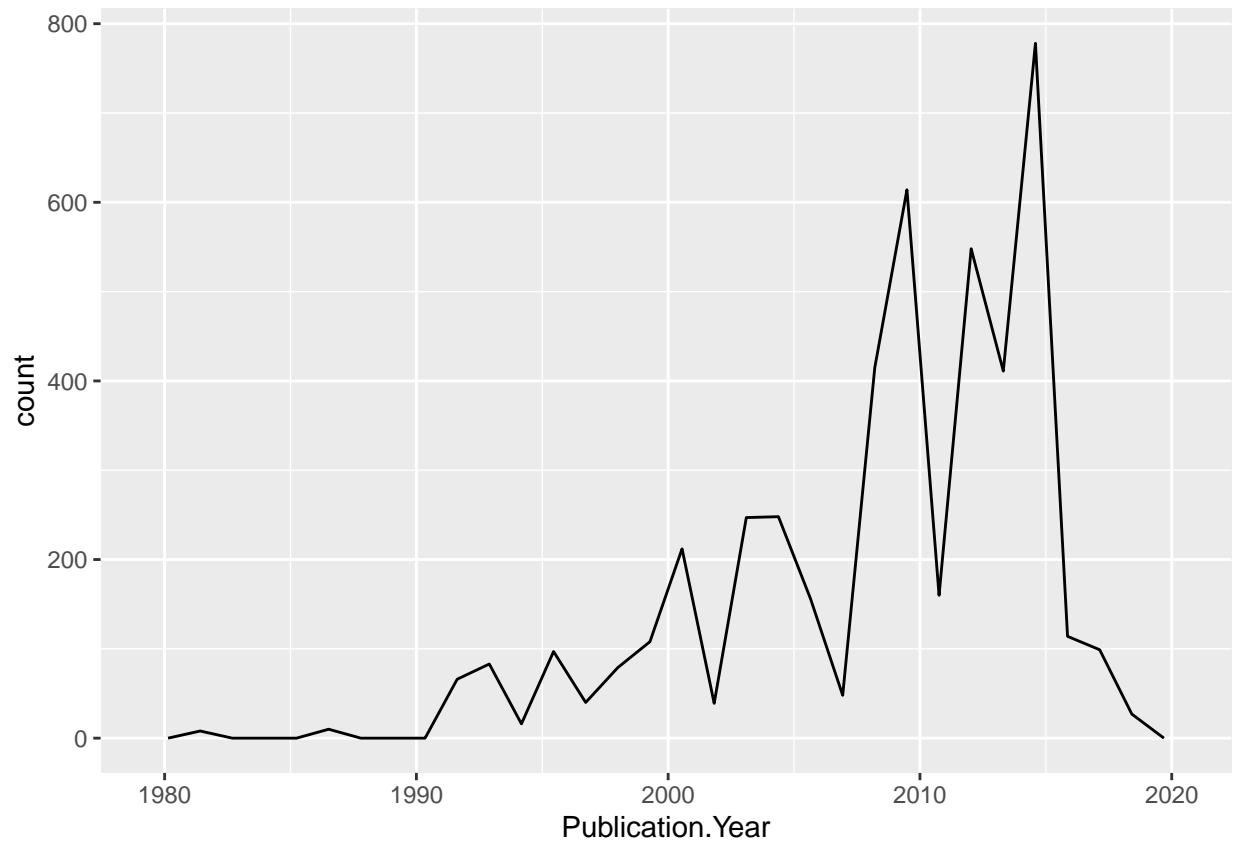
Answer: The field is of type “factor” because of some records listed as “NR” or not recorded.

## Explore your data graphically (Neonics)

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

```
#Create frequency polygon plot of publications by year
ggplot(data = Neonics) +
  geom_freqpoly(aes(x=Publication.Year))
```

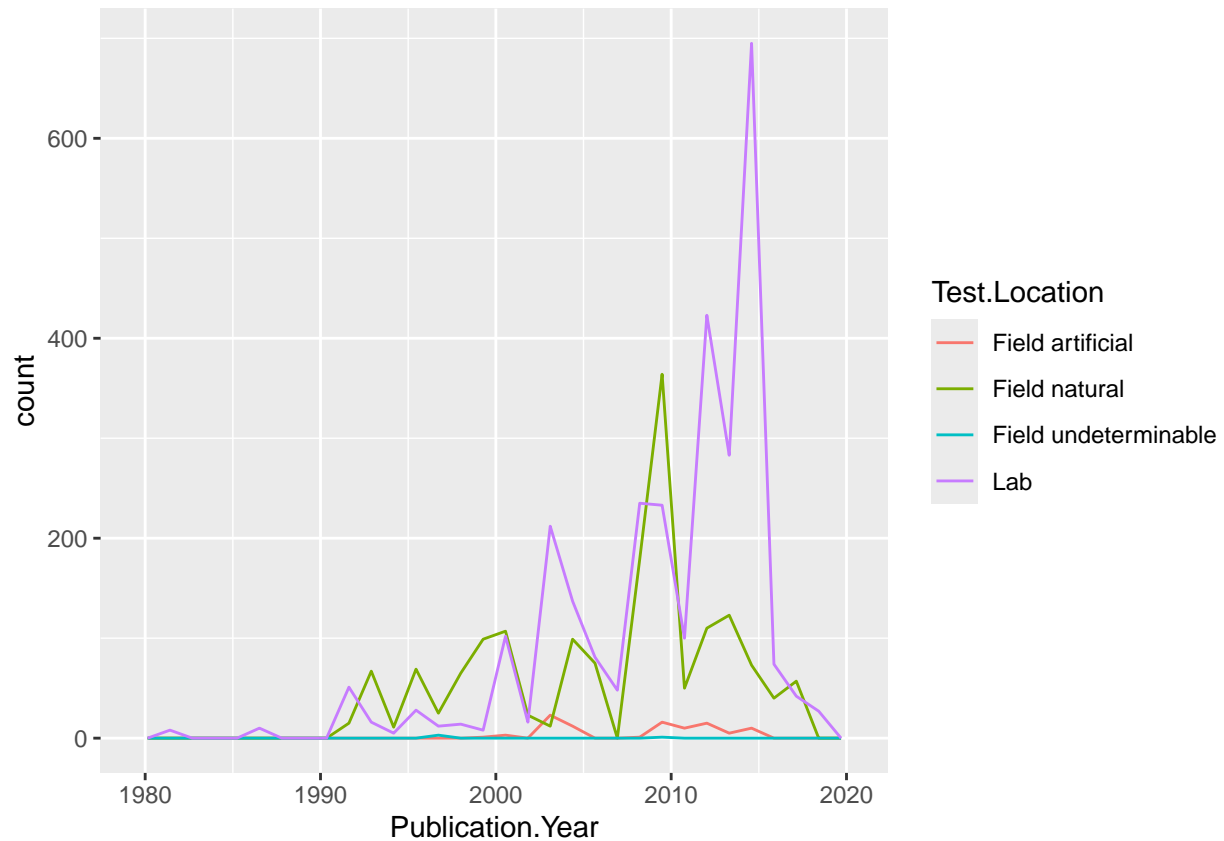
```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



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

```
#Create frequency polygon plot of publications by year, colored by location
ggplot(data = Neonics) +
  geom_freqpoly(aes(x=Publication.Year,color=Test.Location))
```

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



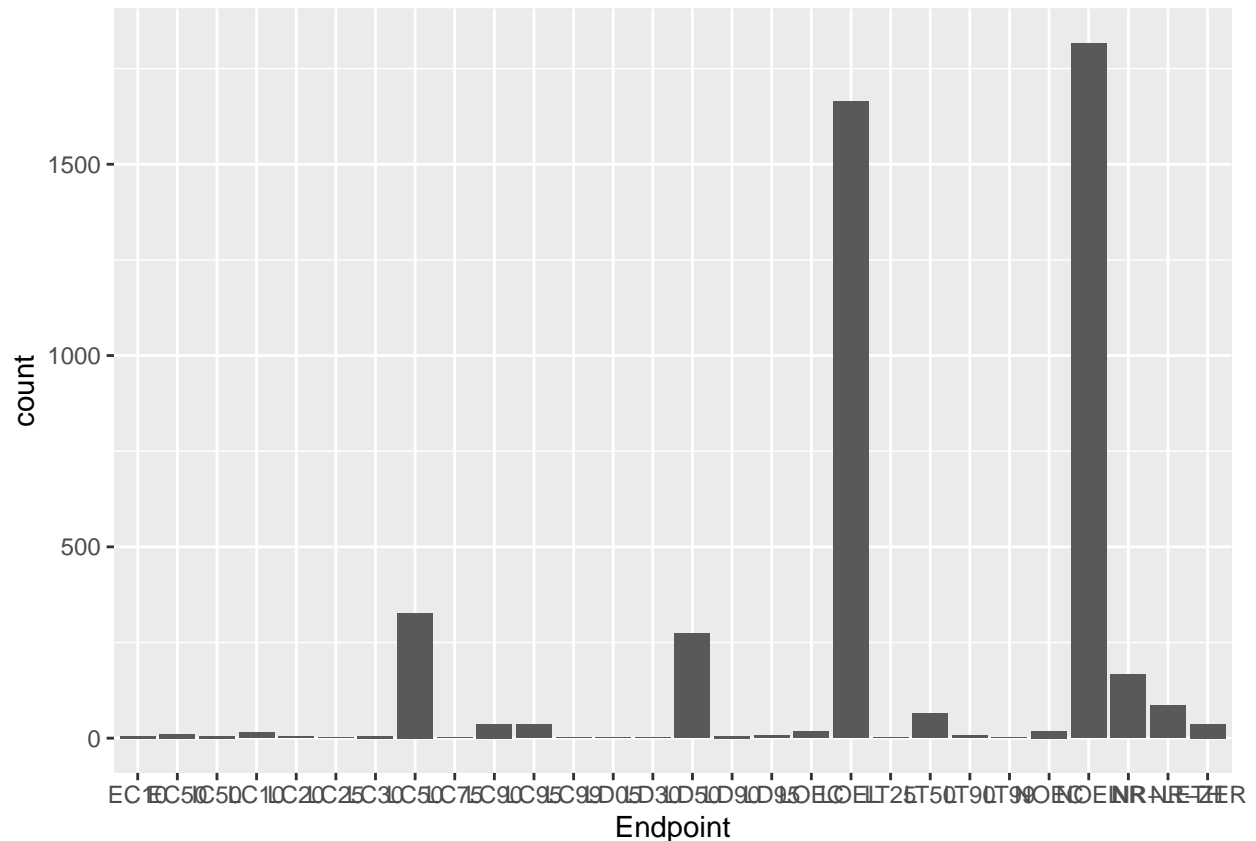
Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: The lab is overall the most common test location, but not always, with Field natural sometimes overtaking it.

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 Endpoint counts
ggplot(Neonics) +
  geom_bar(aes(x=Endpoint))
```



Answer: “NOEL” or No observable effect level is the most common effect, with LOEL or lowest observable effect level the 2nd most common. These are defined as: \* NOEL - No-observable-effect-level: highest dose (concentration) producing effects not significantly different from responses of controls according to author’s reported statistical test \* LOEL - Lowest Observed Effects Residue: The lowest residue concentration producing effects that were significantly different from responses of controls according to author’s reported statistical test

## 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.

```
#Show class of collectDate field
class(Litter$collectDate) #It's a factor, not a date
```

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

```
#Convert to a date
Litter$collectDate <- as.Date(Litter$collectDate,format='%Y-%m-%d')
#Confirm change
class(Litter$collectDate)
```

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



13. Using the `unique` function, determine how many different plots were sampled at Niwot Ridge. How is the information obtained from `unique` different from that obtained from `summary`?

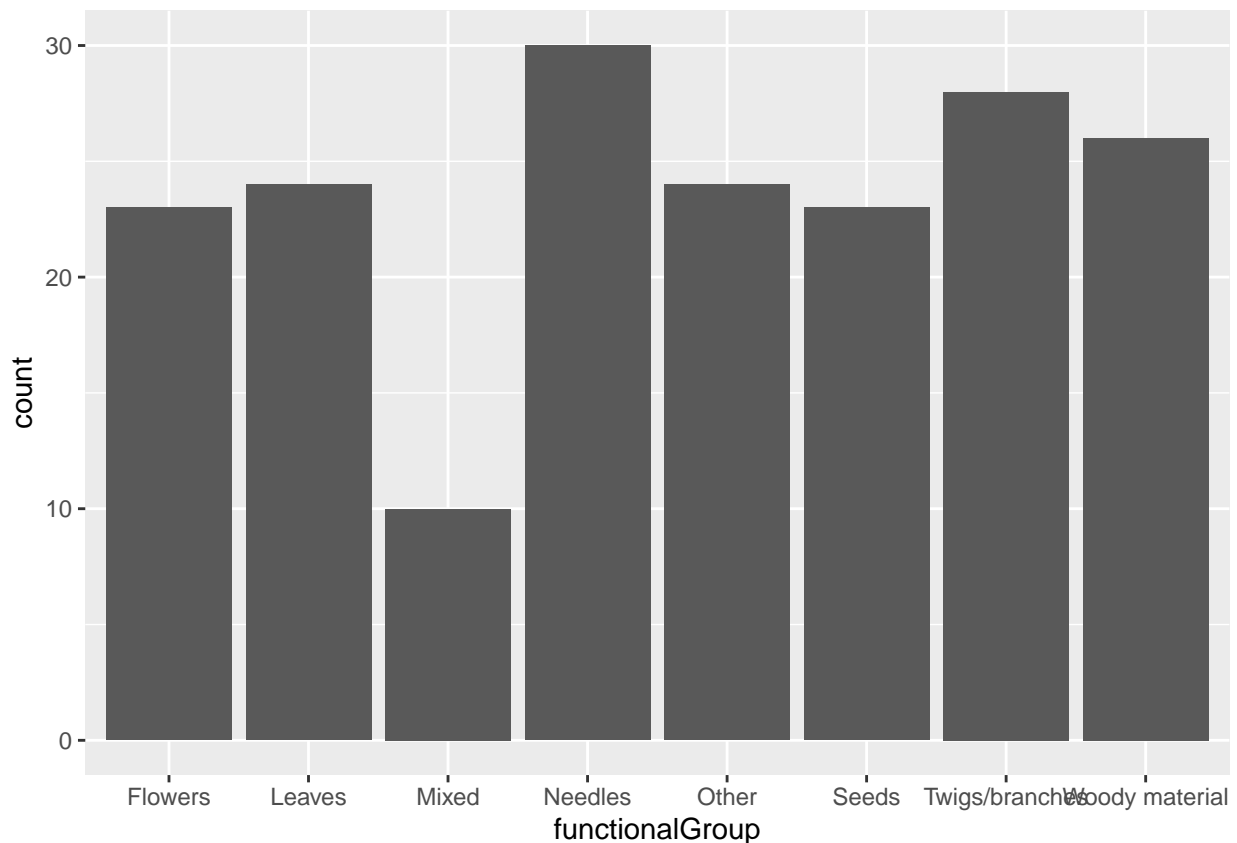
```
#List the unique dates  
unique(Litter$plotID)
```

```
## [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 only lists the unique dates in the study. Summarize will include how many records associate with each.

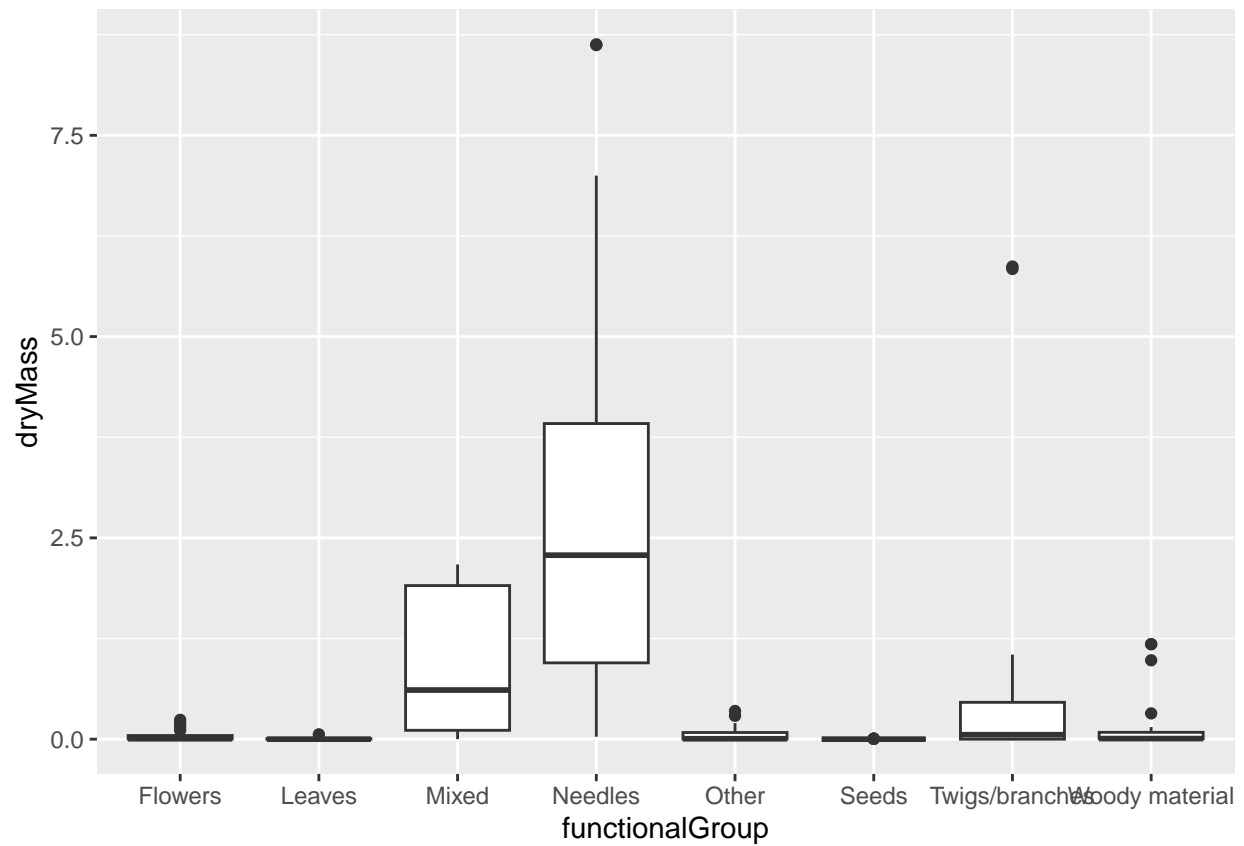
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#Create bar graph of functional group counts  
ggplot(Litter) +  
  geom_bar(aes(x=functionalGroup))
```

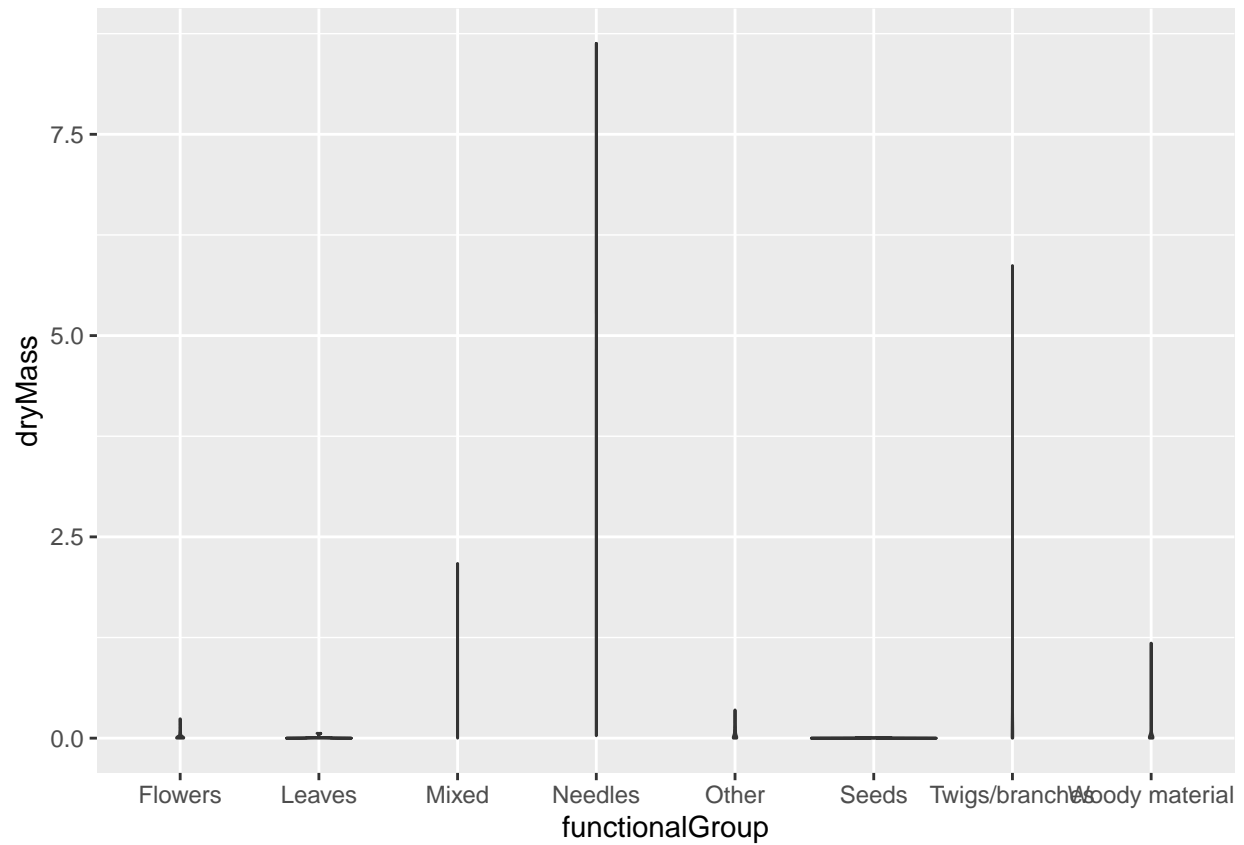


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of `dryMass` by `functionalGroup`.

```
#Create a box plot of dry mass across functional groups
ggplot(Litter) +
  geom_boxplot(aes(y=dryMass,x=functionalGroup))
```



```
#Create a violin plot of dry mass across functional groups
ggplot(Litter) +
  geom_violin(aes(y=dryMass,x=functionalGroup))
```



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

Answer: Not enough values across the distribution of values to reveal densities in a violin plot.

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

Answer: Needles has the highest biomass