

# SamanthaSedar\_A03\_DataExploration.Rmd

Samantha Sedar

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.

```
getwd()
```

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

```
library(tidyverse)
library(lubridate)
```

```
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = T)
Litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv", stringsAsFactors = T)
```

## 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: According to Hladik, Main, and Goulson, neonicotinoids are known to have adverse impacts on pollinators and negatively impact aquatic insects and ecosystems. Given the importance of pollinators to our ecosystems and the environment more broadly, it is prudent to study the ecotoxicology of neonicotinoids on insects. Source: *nvirion. Sci. Technol.* 2018, 52, 6, 3329–3335  
Publication Date: February 26, 2018 <https://doi.org/10.1021/acs.est.7b06388> y.

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: According to Scheungrab, Trettin, Lea, and Jurgensen, woody debris is important to study due to its role in carbon budgets and nutrient cycling in addition to influencing water flows and sediment transport. Source: In: *Gen. Tech. Rep. SRS-38*. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. p. 47-48.

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. Spatial sampling design: Tower plots are randomly selected within the 90% flux footprint of primary/secondary airsheds 2. Spatial sampling design: Plot edges must be separated by a distance 150% of one edge of the plot 3. Temporal sampling design: Sampling frequency varies between biweekly and bimonthly depending on vegetation while ground traps are sampled annually

## Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
print(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?

```
sort(summary(Neonics$Effect))
```

##	Hormone(s)	Histology	Physiology	Cell(s)
##	1	5	7	9
##	Biochemistry	Accumulation	Intoxication	Immunological
##	11	12	12	16
##	Morphology	Growth	Enzyme(s)	Genetics
##	22	38	62	82
##	Avoidance	Development	Reproduction	Feeding behavior
##	102	136	197	255
##	Behavior	Mortality	Population	
##	360	1493	1803	

Answer: The most common effects that are studied are by far population and mortality at 1803 and 1493, respectively. These effects are likely the most commonly studied because they are the most definite/important to overall research, they are tightly related to each other and also each of the other effects.

- 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...]

```
sort(summary(Neonics$Species.Common.Name))
```

##	Ant Family	Apple Maggot
##	9	9
##	Glasshouse Potato Wasp	Lacewing
##	10	10
##	Southern House Mosquito	Two Spotted Lady Beetle
##	10	10
##	Spotless Ladybird Beetle	Braconid Parasitoid
##	11	12
##	Common Thrip	Eastern Subterranean Termite
##	12	12
##	Jassid	Mite Order
##	12	12
##	Pea Aphid	Pond Wolf Spider
##	12	12
##	Armoured Scale Family	Diamondback Moth
##	13	13
##	Eulophid Wasp	Monarch Butterfly
##	13	13
##	Predatory Bug	Yellow Fever Mosquito
##	13	13
##	Corn Earworm	Green Peach Aphid
##	14	14
##	House Fly	Ox Beetle
##	14	14
##	Red Scale Parasite	Spined Soldier Bug
##	14	14
##	Western Flower Thrips	Hemlock Woolly Adelgid Lady Beetle
##	15	16
##	Hemlock Woolly Adelgid	Mite
##	16	16

##	Onion Thrip	Araneoid Spider Order
##	16	17
##	Bee Order	Egg Parasitoid
##	17	17
##	Insect Class	Moth And Butterfly Order
##	17	17
##	Oystershell Scale Parasitoid	Black-spotted Lady Beetle
##	17	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
##	Codling Moth	Flatheaded Appletree Borer
##	19	20
##	Horned Oak Gall Wasp	Leaf Beetle Family
##	20	20
##	Potato Leafhopper	Tooth-necked Fungus Beetle
##	20	20
##	Argentine Ant	Beetle
##	21	21
##	Mason Bee	Mosquito
##	22	22
##	Citrus Leafminer	Ladybird Beetle
##	23	23
##	Spider/Mite Class	Tobacco Flea Beetle
##	24	24
##	Chalcid Wasp	Convergent Lady Beetle
##	25	25
##	Stingless Bee	Ground Beetle Family
##	25	27
##	Rove Beetle Family	Tobacco Aphid
##	27	27
##	Scarab Beetle	Spring Tiphia
##	29	29
##	Thrip Order	Ladybird Beetle Family
##	29	30
##	Parasitoid	Braconid Wasp
##	30	33
##	Cotton Aphid	Predatory Mite
##	33	33
##	Sweetpotato Whitefly	Aphid Family
##	37	38
##	Cabbage Looper	Buff-tailed Bumblebee
##	38	39
##	True Bug Order	Sevenspotted Lady Beetle
##	45	46
##	Beetle Order	Snout Beetle Family, Weevil
##	47	47
##	Erythrina Gall Wasp	Parasitoid Wasp
##	49	51

##	Colorado Potato Beetle	Parastic Wasp
##	57	58
##	Asian Citrus Psyllid	Minute Pirate Bug
##	60	62
##	European Dark Bee	Wireworm
##	66	69
##	Euonymus Scale	Asian Lady Beetle
##	75	76
##	Japanese Beetle	Italian Honeybee
##	94	113
##	Bumble Bee	Carniolan Honey Bee
##	140	152
##	Buff Tailed Bumblebee	Parasitic Wasp
##	183	285
##	Honey Bee	(Other)
##	667	670

Answer: The six most commonly studied species are: 1. Honey Bee-667; 2. Parasitic Wasp-285; 3. Buff Tailed Bumblebee-183; 4. Carniolan Honey Bee-152; 5. Bumble Bee-140; 6. Italian Honeybee-113. The commonality here is that they are all bees, and as indicated in answer #2, neonicotinoids are known to have adverse impacts on pollinators, so it is in-line with expectations that bees (primary pollinators) would be the insects of most interest.

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.Units..Author.)
```

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

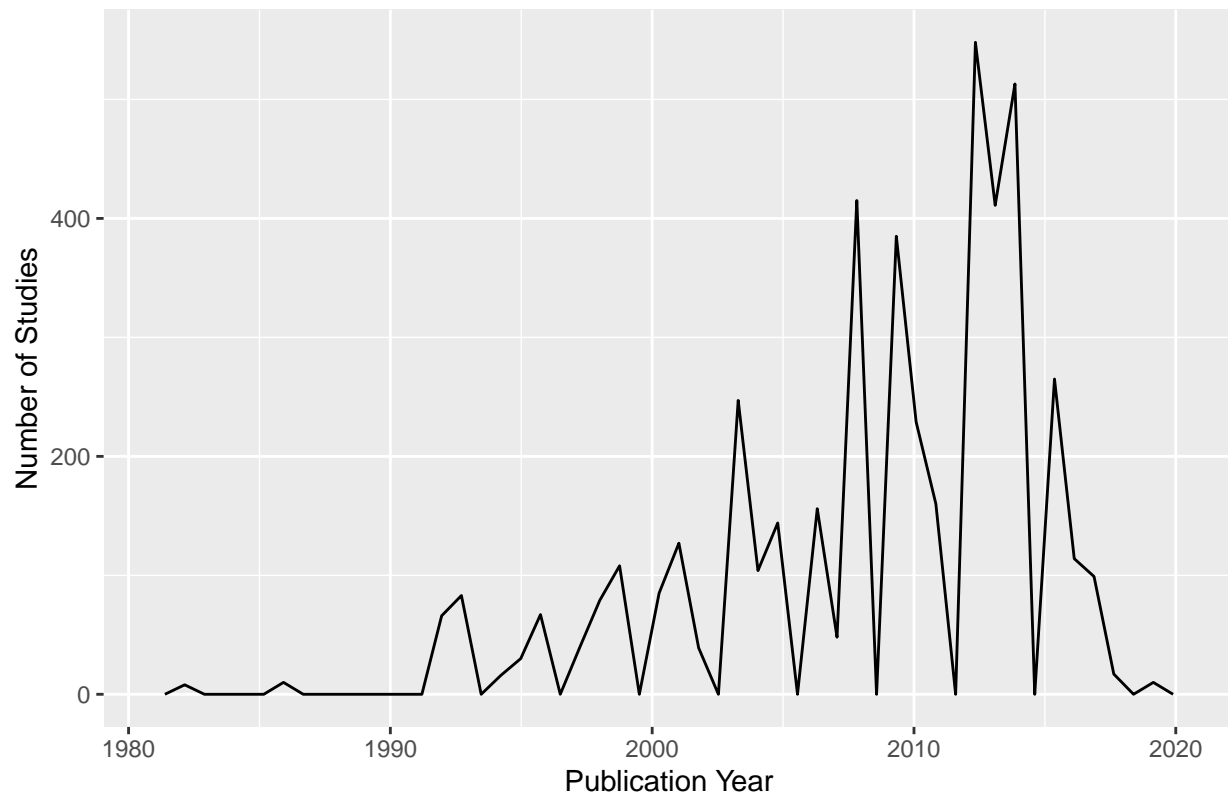
Answer: The concentration records, 'Conc.1..Author.' is a factor and not numeric because the column contains several non-numeric values. In order to convert an entire column to numeric, each value would need to have a numeric value.

## Explore your data graphically (Neonics)

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

```
ggplot(Neonics) +
  geom_freqpoly(aes(x= Publication.Year), bins=50) +
  labs(title = "Number of Studies by Publication Year",
       x= "Publication Year",
       y = "Number of Studies")
```

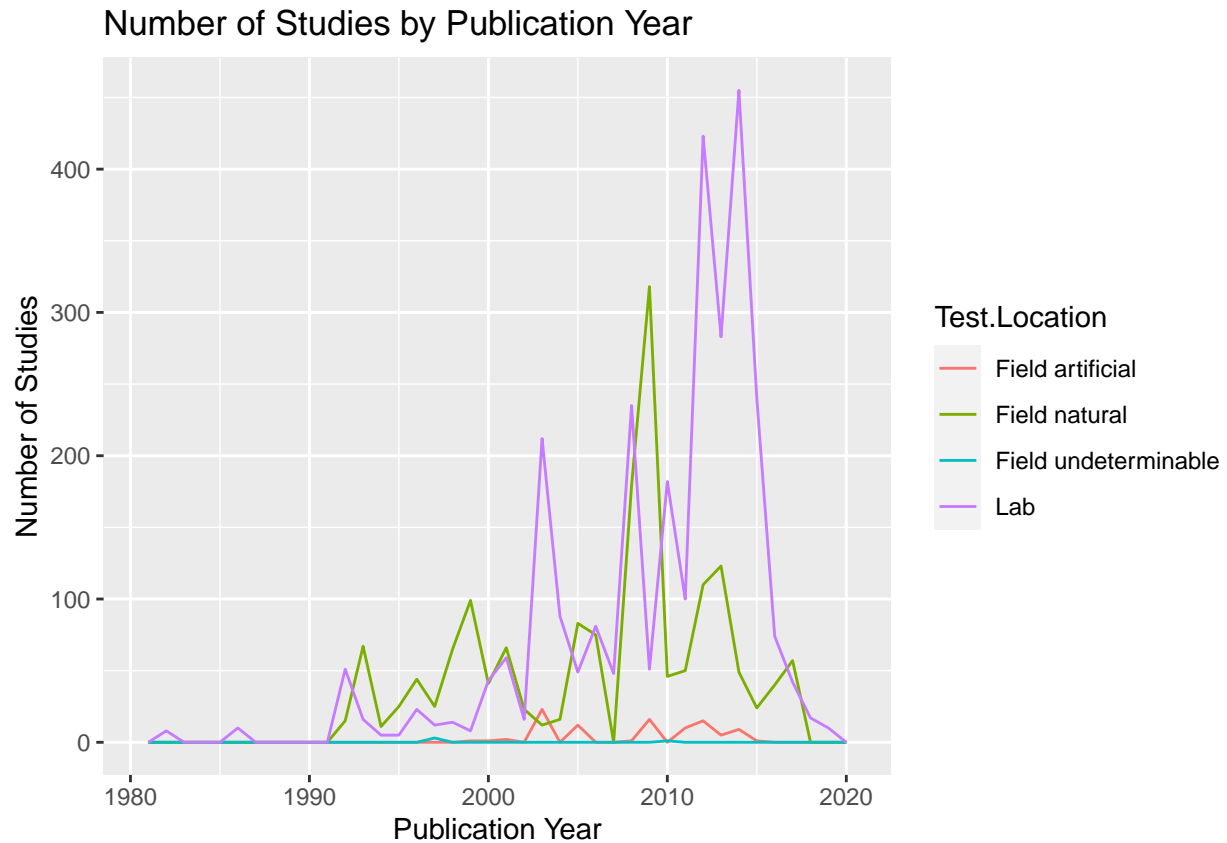
Number of Studies by Publication Year



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

#standard syntax + adding color=test.location to make the locations display as different colors

```
ggplot(Neonics, aes(x=Publication.Year, color = Test.Location)) +  
  geom_freqpoly(binwidth=1) +  
  labs(title = "Number of Studies by Publication Year",  
        x= "Publication Year",  
        y = "Number of Studies")
```



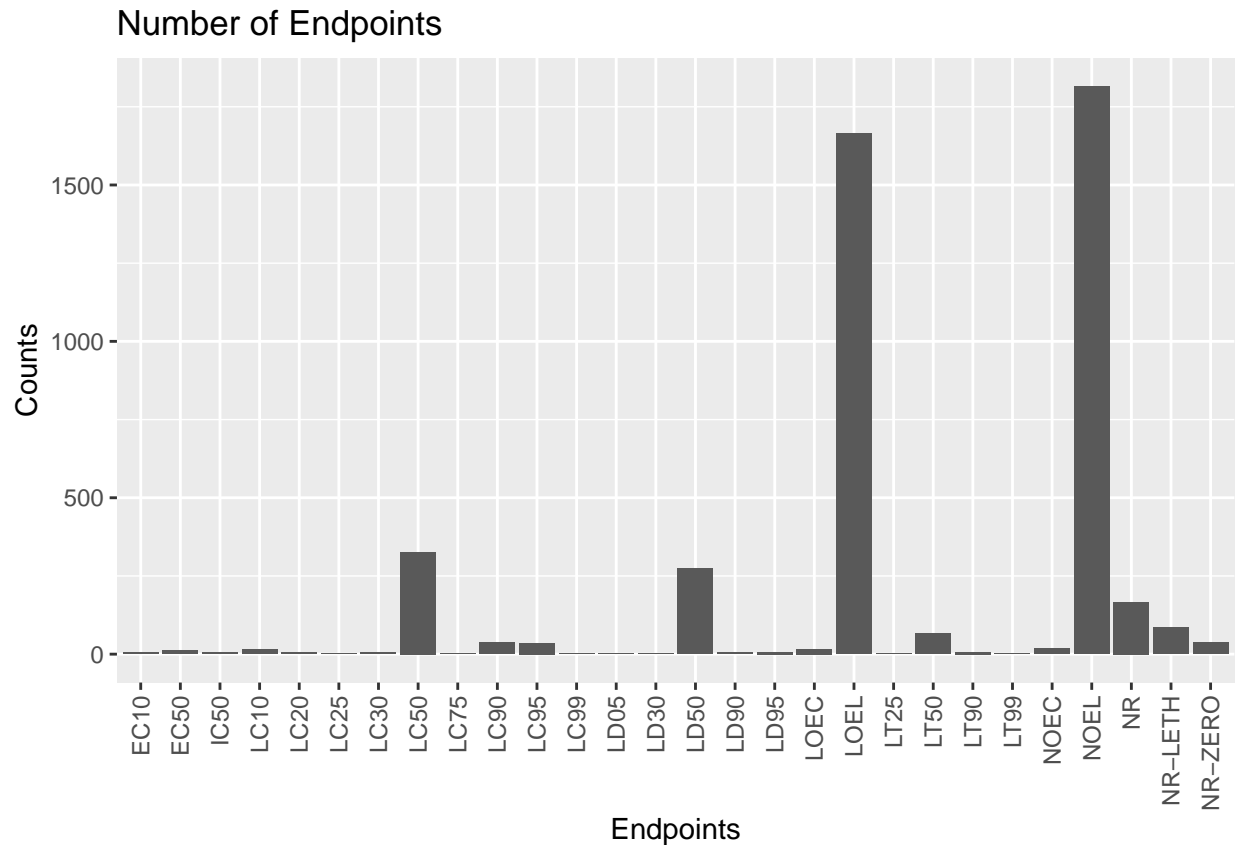
Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: The most common test locations differ over time. However, based on this graph, we can see that 'field artificial' is the least common and the most common is 'lab.' This is likely due to the accessibility and reliability of a lab, as compared to the remaining options.

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)) +
  labs(title = "Number of Endpoints",
       x = "Endpoints",
       y = "Counts")
```



Answer: The two most common endpoints are: 1. NOEL-Terrestrial: No-observable-effect-level: highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test (NOEL/NOEC) 2. LOEL-Terrestrial: Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEL/LOEC)

## 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)
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
class(Litter$collectDate)
```

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

```
unique_dates_aug_2018 <- unique(Litter$collectDate[format(Litter$collectDate, "%Y-%m") == "2018-08"])
print(unique_dates_aug_2018)
```



```
## [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`?

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

```
## [1] 12
```

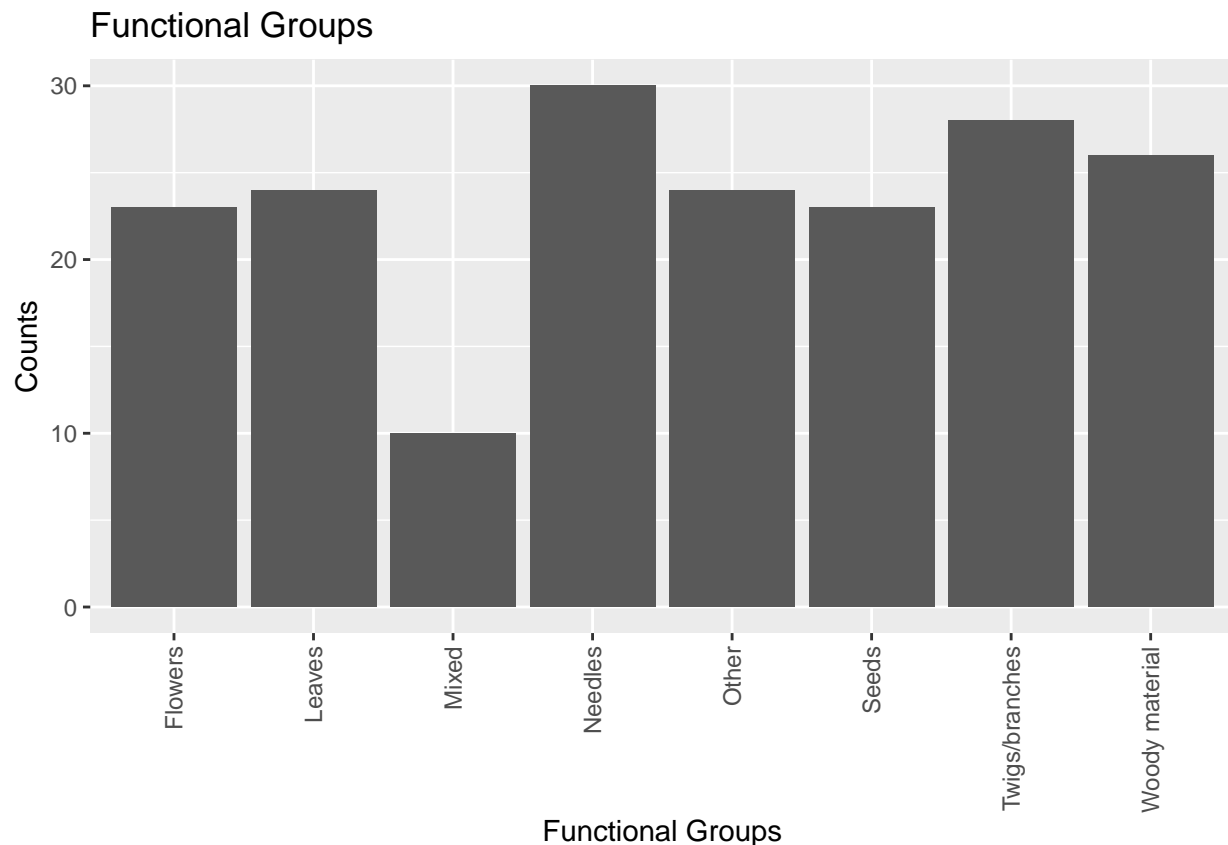
```
length(summary(Litter$plotID))
```

```
## [1] 12
```

Answer: The information obtained using the `unique` function along with the `length` function result in the same information obtained from using `summary`. This is because the `plotID` is a factor, therefore there are no other summary statistics that can be generated.

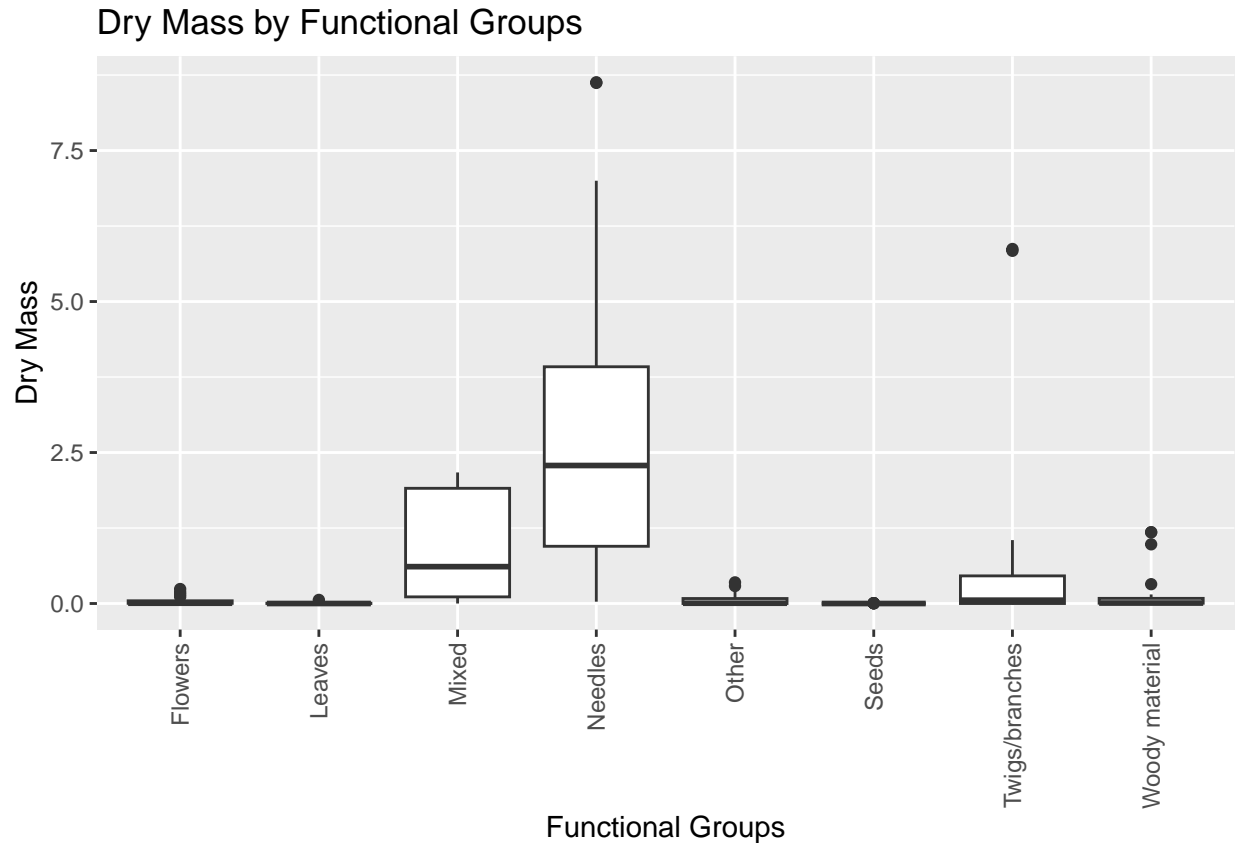
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(x = functionalGroup)) +  
  geom_bar() + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +  
  labs(title = "Functional Groups",  
        x = "Functional Groups",  
        y = "Counts")
```



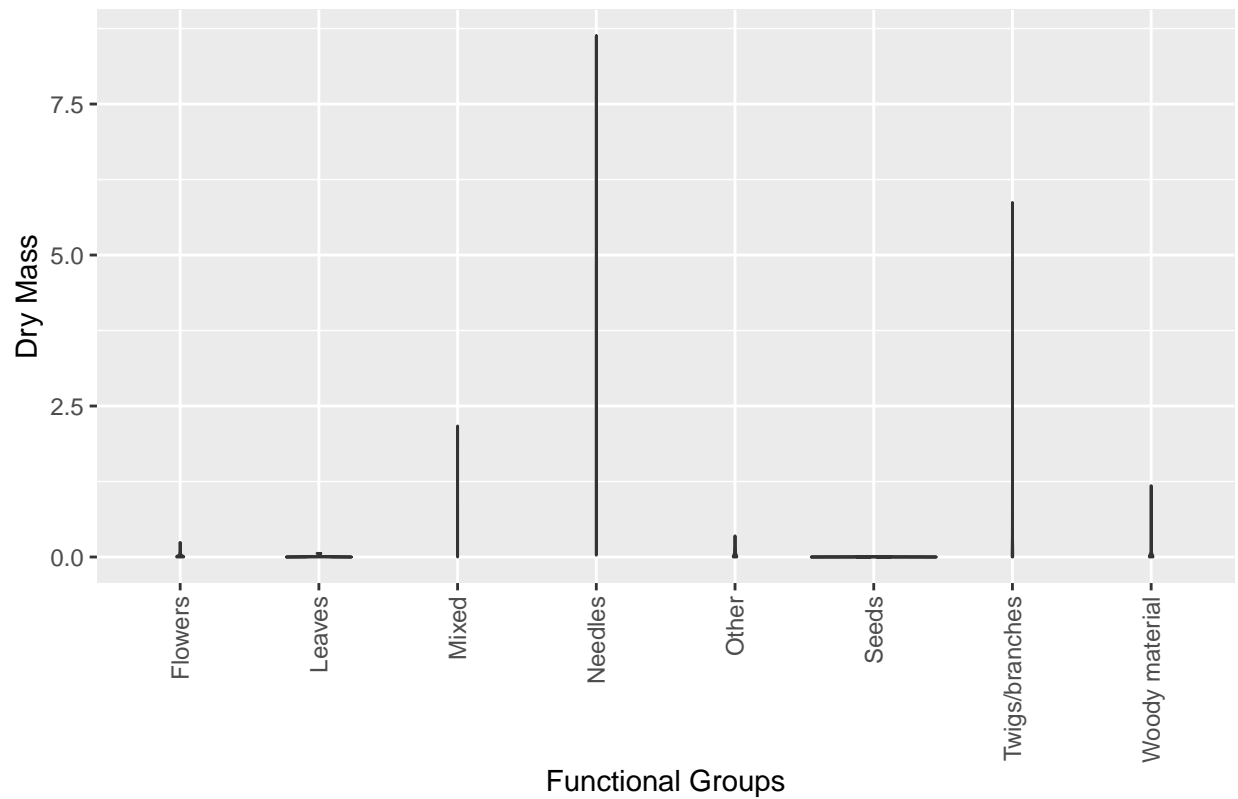
15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of `dryMass` by functional-Group.

```
ggplot(Litter) +  
  geom_boxplot(aes(x = functionalGroup, y = dryMass)) + theme(  
    axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +  
  labs(title = "Dry Mass by Functional Groups",  
        x = "Functional Groups",  
        y = "Dry Mass")
```



```
ggplot(Litter) +  
  geom_violin(aes(x = functionalGroup, y = dryMass)) + theme(  
    axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +  
  labs(title = "Dry Mass by Functional Groups",  
        x = "Functional Groups",  
        y = "Dry Mass")
```

# Dry Mass by Functional Groups



```
length(summary(Litter$Flowers))
```

```
## [1] 3
```

```
length(summary(Litter$Leaves))
```

```
## [1] 3
```

```
length(summary(Litter$Mixed))
```

```
## [1] 3
```

```
length(summary(Litter$Needles))
```

```
## [1] 3
```

```
length(summary(Litter$Other))
```

```
## [1] 3
```

```
length(summary(Litter$Seeds))
```

```
## [1] 3
```

```
length(summary(Litter$"Twings/branches"))
```

```
## [1] 3
```

```
length(summary(Litter$"Woody material"))
```

```
## [1] 3
```

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

Answer: A violin plot is an effective way to show density distributions. For this data, the boxplot is more effective than the violin plot because there is not enough data to depict a trend. Indeed, each functional group only has three datapoints.

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.