

Assignment 3: Data Exploration

Aman Hingu, Section #1

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)
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: The ecotoxicology of neonicotinoids is of great importance for environmentalists and researchers. Insects represent a large group of organisms that play a variety of ecological roles in maintaining stable predator prey relationships, nutrient levels (nitrogen, phosphorus, etc.), and can even assist in the pollination of various plants. Thus, the long term and short term effects of neonicotinoids on non-targeted insect populations is of major concern and should be managed effectively.

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: Litter and woody debris that deposit on forest grounds can each have unique ways of degrading in the environment. Some materials can degrade biologically with the assistance of soil microbiota, but many cannot. Thus, it is important to track woody debris and litter so researchers can track persistence of potentially non degrading materials deposited. Nondegradable materials can pose environmental health risks to species in the ecosystem.

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 traps are placed in pairs, with one being an elevated trap and the other being a ground trap. One litter trap pair is placed in each 400 square meter area. * Litter traps are placed strategically, or randomly in areas with greater than 50% aerial cover of woody vegetation greater than two meters in height. * The frequency of sampling performed on elevated traps can vary depending on the density of vegetation present, but ground traps are sampled once a year.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the 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?

```
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: Based on the calculated summary of the Effects column, it appears that the most common effects studied are population, mortality, behavior, and feeding behavior. Looking at the population and mortality parameters would give researchers an easy indicator of the more direct effects of various compounds. Behavior and feeding behavior can provide researchers an insight into less direct sensitivities of various organisms as a result of exposure to the compounds studied.

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 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: The six most commonly studied insects in this study are the Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and the Italian Honeybee. These bees and wasps are most likely being studied due to their ecological importance to biodiversity and ecosystem stability. For example, bees are some of the most important pollinators of many fruits, vegetables, and flowers. This has a direct impact on other species in the ecosystems they dwell in, but also food supply and agriculture.

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 factor. R does not recognize this data as numeric due to the presence of slash's in some of the rows.

Explore your data graphically (Neonics)

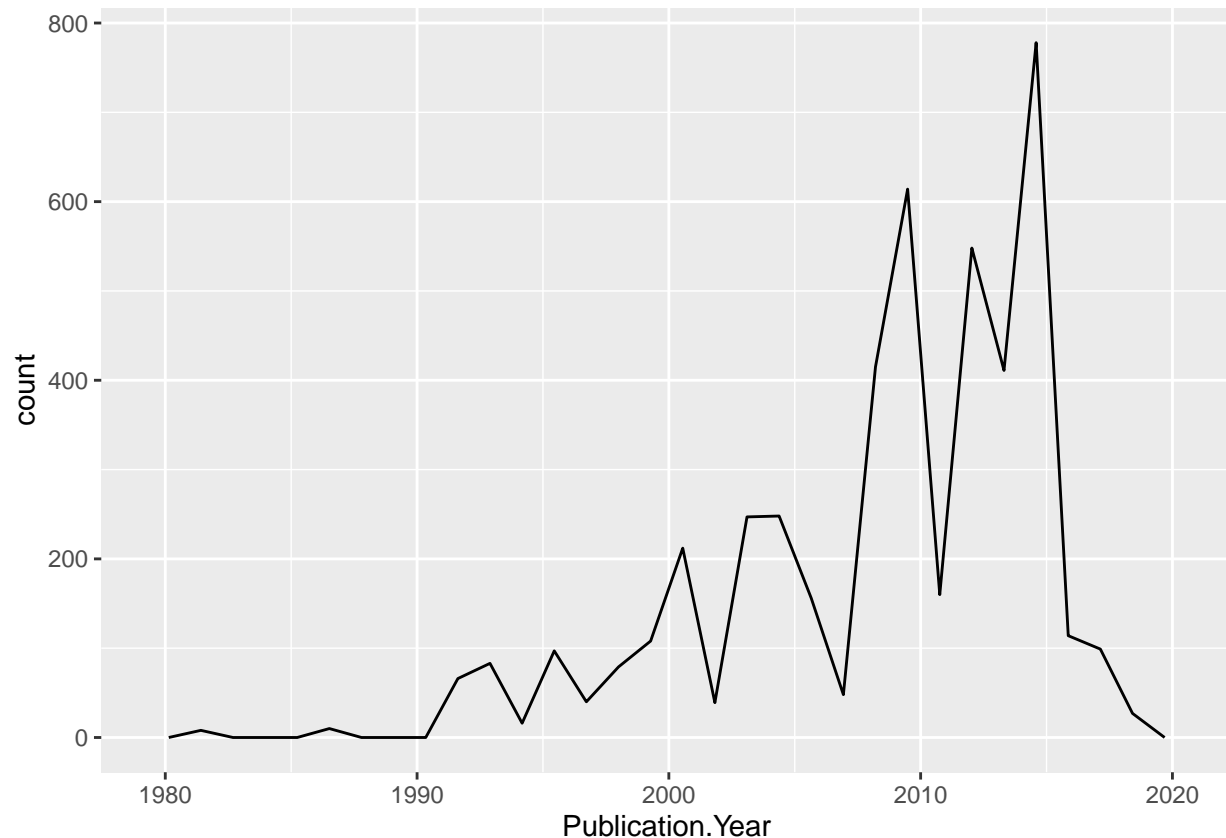
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
str(Neonics$Publication.Year)
```

```
## int [1:4623] 1982 1982 1986 1986 1986 1986 1986 1986 1986 1986 ...
```

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

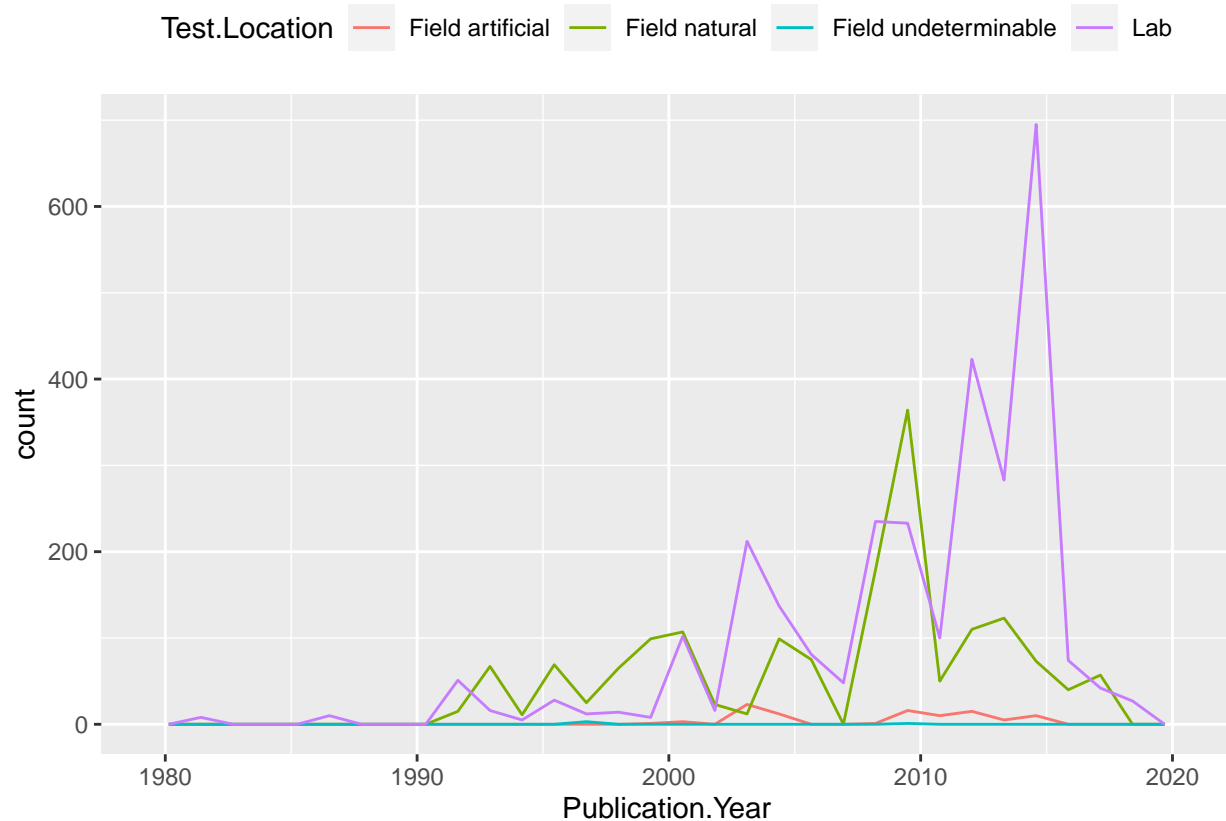
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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)) +  
  theme(legend.position = "top")
```

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

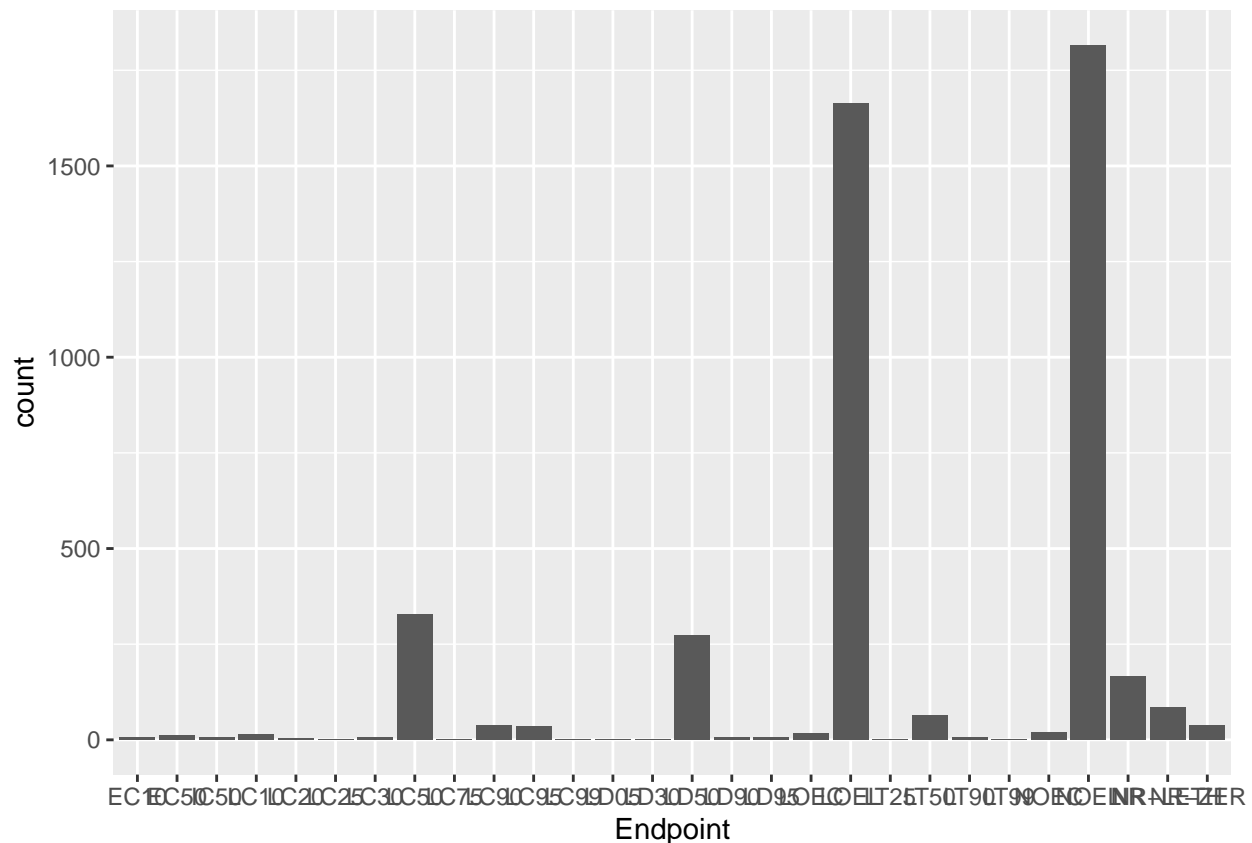


Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: According to the graph above, it appears that the most common test locations are the “Lab” and “Field Natural” sites. Over time, there is a net increase in the occurrence of the “Lab” test location as compared to the “Field Natural” test location. This is most clearly observed from 2010-2020.

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, aes(x = Endpoint)) +  
  geom_bar()
```



Answer: According to the bar graph above, the two most commonly occurring Endpoints in the study are the No Observed Effect Level (NOEL) and the Lowest Observed Effect Level (LOEL). According to the ECOTOX Code Appendix, the LOEL is defined as the lowest dose producing effects that were significantly different (as reported by authors) from responses of controls, and the NOEL was defined as the highest dose producing effects not significantly different from the responses of controls according to the 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.

```
class(Litter$collectDate)
```

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

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

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

```
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`?

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

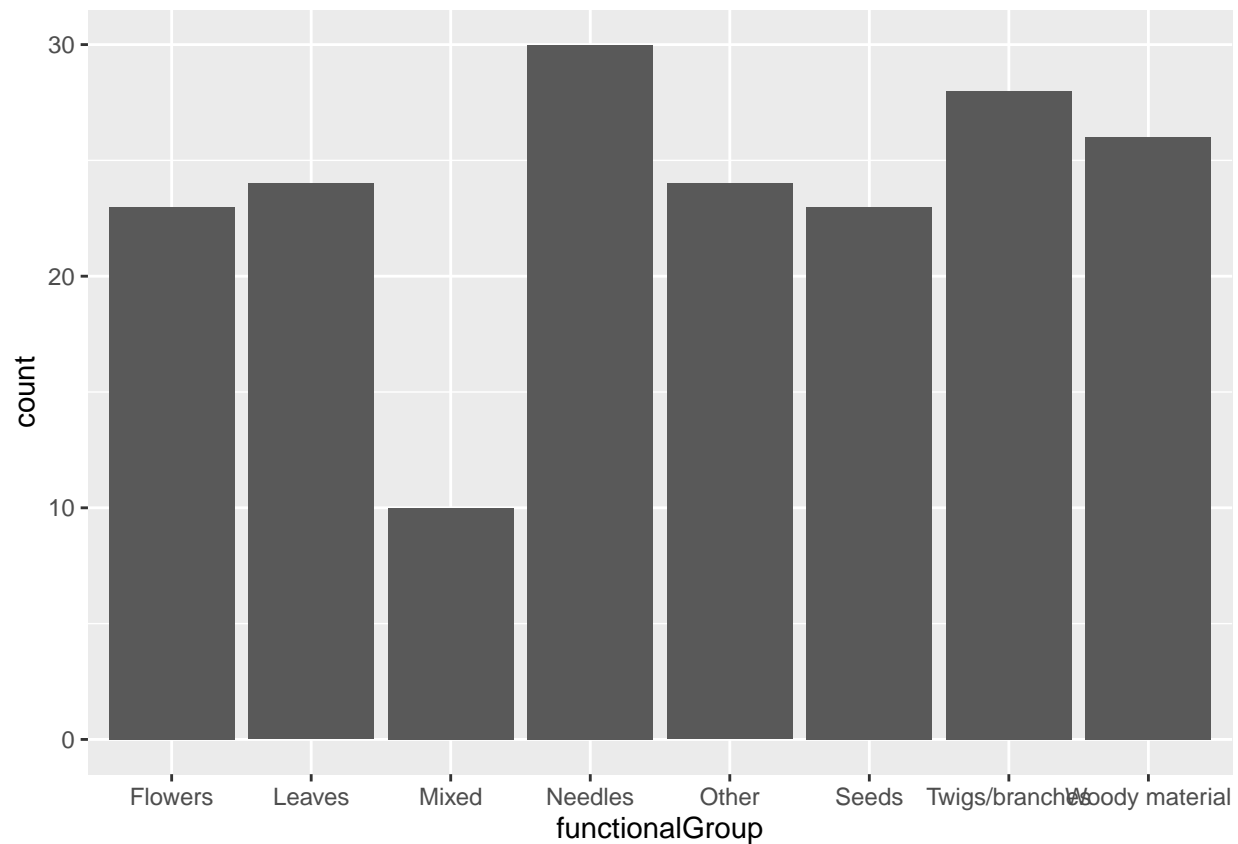
```
## [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051
```

```
## [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: According to the output above, 12 plots were sampled at Niwot Ridge. The information obtained here differs from the information provided by the summary function because the unique() function provides all the different & unique values obtained for a certain parameter.

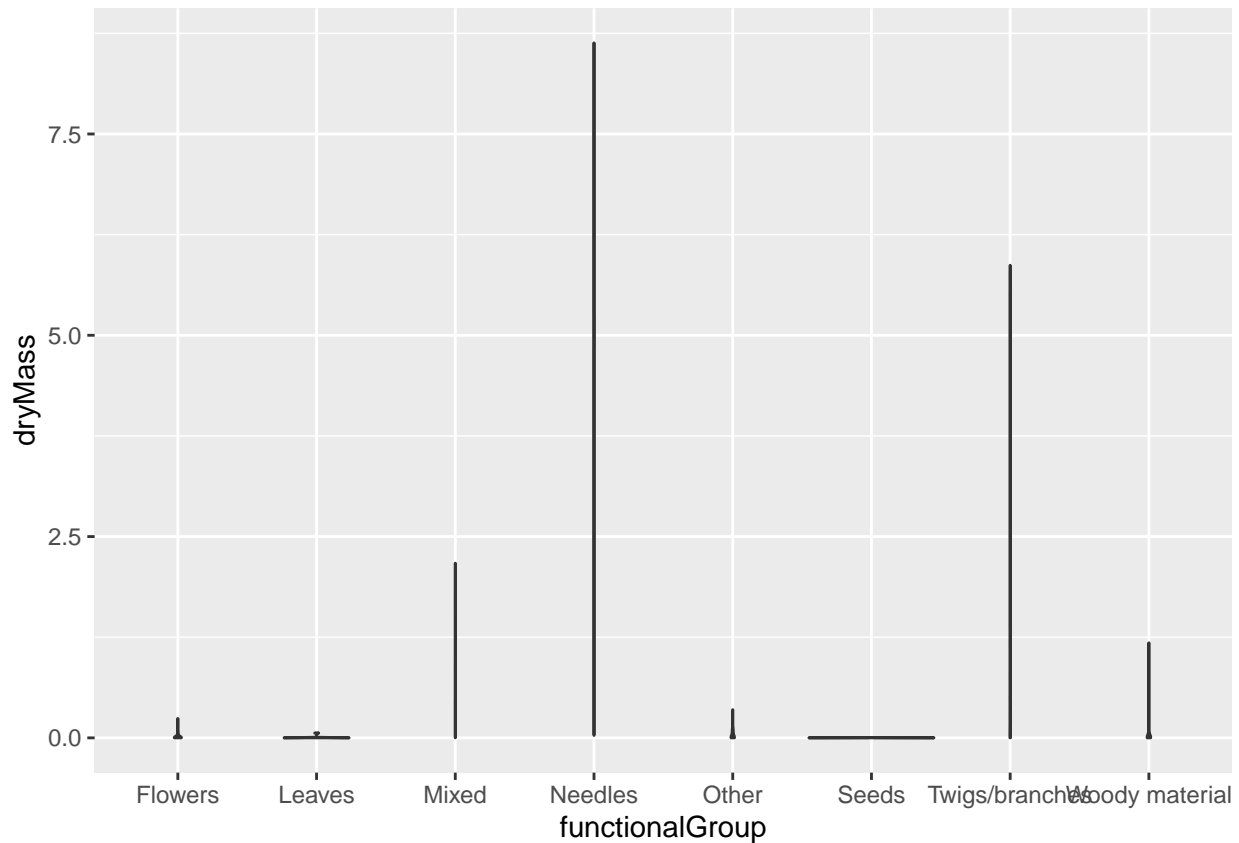
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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



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

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

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

Answer: A boxplot is a more effective visualization option as compared to the violin plot in this case because it can indicate any potential skewness within each group with the clearly displayed median.

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

Answer: Of all the various types of litter present, the needles and mixed groups constitute the highest levels of biomass.