

Assignment 3: Data Exploration

Shiqi Zheng

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.

```
# preparation  
getwd()
```

```
## [1] "/Users/shiqizheng/Desktop/ENV872/EDE_Fall2023"
```

```
library(tidyverse)  
library(lubridate)  
# load data  
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: Neonicotinoid is toxic for a broad range of insects. It can harm non-target organisms like pollinators, and it can accumulate in insects, potentially affecting predators and disrupting the food chain. By altering insect population, it can disrupt ecosystems. (References: Hladik, M. L., Main, A. R., & Goulson, D. (2018). Environmental risks and challenges associated with neonicotinoid insecticides.)

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: Woody debris plays a significant role within forest and stream ecosystems. It contributes to carbon budgets and nutrient cycling, serves as an energy source for aquatic ecosystems, offers habitats for both terrestrial and aquatic organisms, and adds to structural complexity. It can impact water flows and sediment transport through its influence on structure and surface roughness. (References: Scheungrab, Donna B.; Trettin, Carl C.; Lea, Russ; Jurgensen, Martin F. 2000. Woody debris. 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. There are temporal and spatial sampling design. Sampling for this product occurs only in tower plots. Sampling occurs in the same locations over the lifetime of the Observatory if possible. 2. For spatial sampling, in sites with forested tower airsheds, the litter sampling designed 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 26 20m x 20m plots. 3. For temporal sampling, 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
```

Dimension of the dataset is 4623 observations of 30 variables.

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: The most common effects that are studied are population (1803). One of the major concern about neonicotinoids is it affects a broad spectrum of insects. And due to neonicotinoids is widely used as a insecticides, population change can be one of the most direct indicator of the effects of neonicotinoids.

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...]

```
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 in the dataset are: Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee ,Bumble Bee, Italian Honeybee. All these species are belong to the order Hymenoptera, and they play important roles in ecosystem. Bees are vital pollinators, while parastic wasp can influence insect population by parasitizing. They are also important in economic, bees has impacts on food production and parastic wasp can be used to control pests.

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"
```

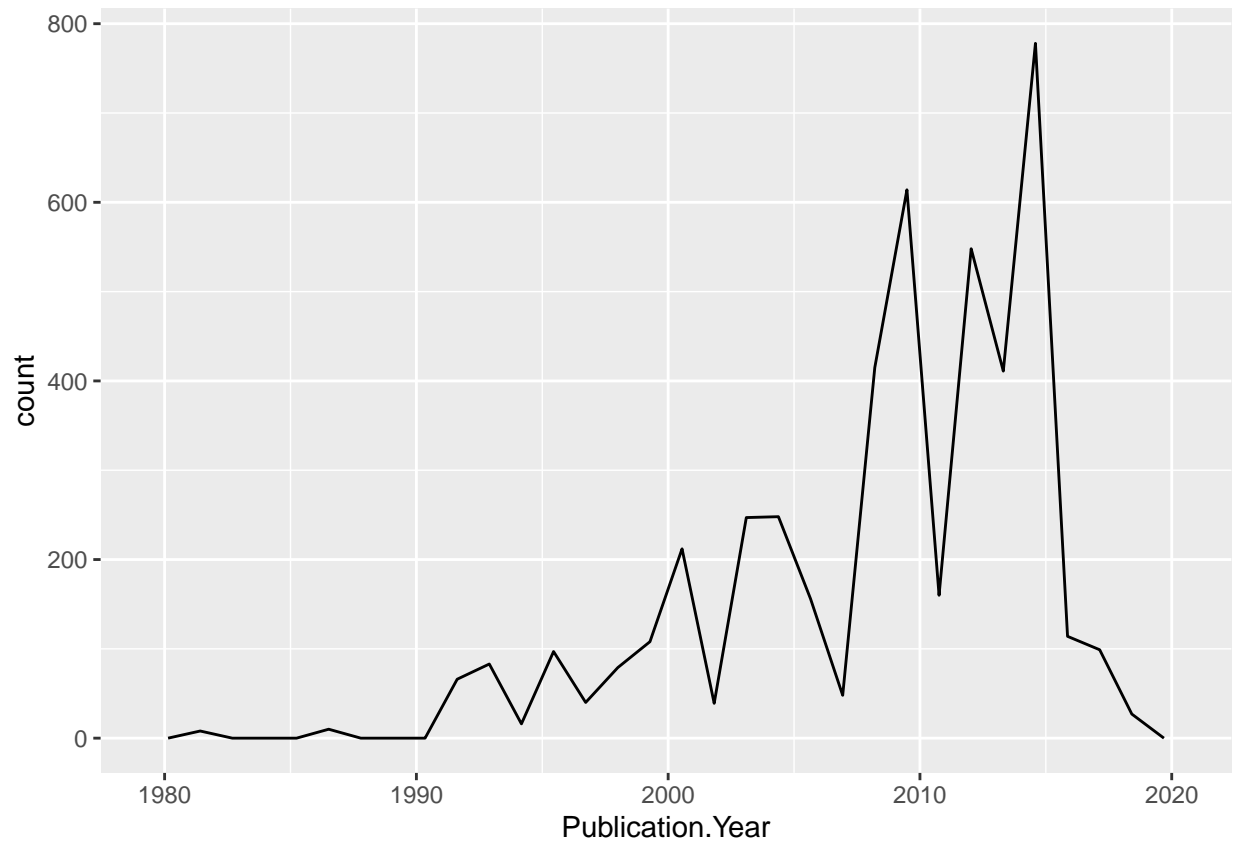
Answer: The class of `Conc.1..Author.` column is factor. It is not numeric because in the dataset it contains value like 'NR', '~10', '144.0/', which are not purely numeric.

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))
```

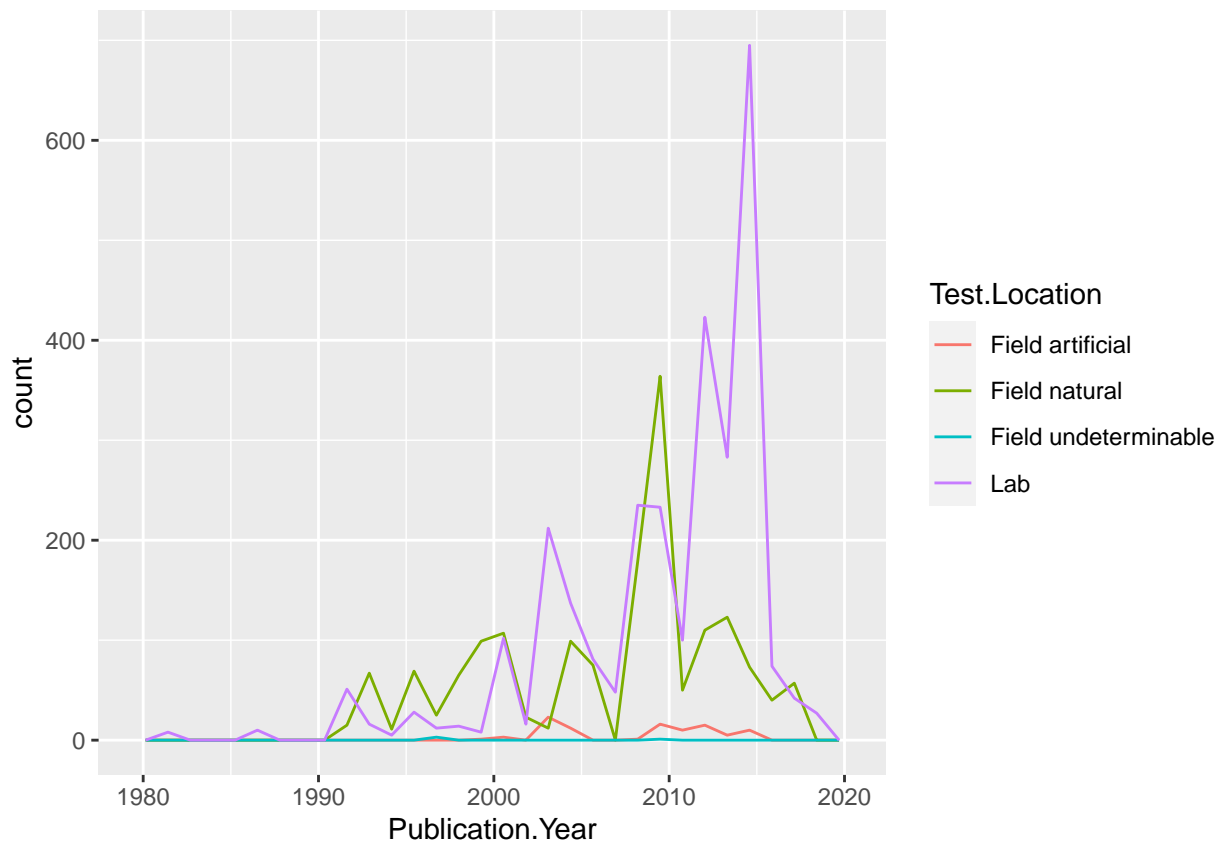
```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



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: Before 2000, field natural is the most common test locations. Between 2000 to ~2010, lab and field natural are alternated to be the most. After ~2010, lab is the most common test locations.

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: LOEL and NOEL are the two most common end points. LOEL used for terrestrial database, represents Lowest-observable-effect-level, meaning lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC). NOEL also used for terrestrial database, represents No-observable-effect-level, which means 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) # factor
```

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

```
# change to date
```

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

```
class(Litter$collectDate) # date
```

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



```
# determine sample date  
unique(x=Litter$collectDate)
```

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

Litter was sampled in 2nd and 30th in August 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(x=Litter$siteID)
```

```
## [1] NIWO  
## Levels: NIWO
```

```
nrow(Litter)
```

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

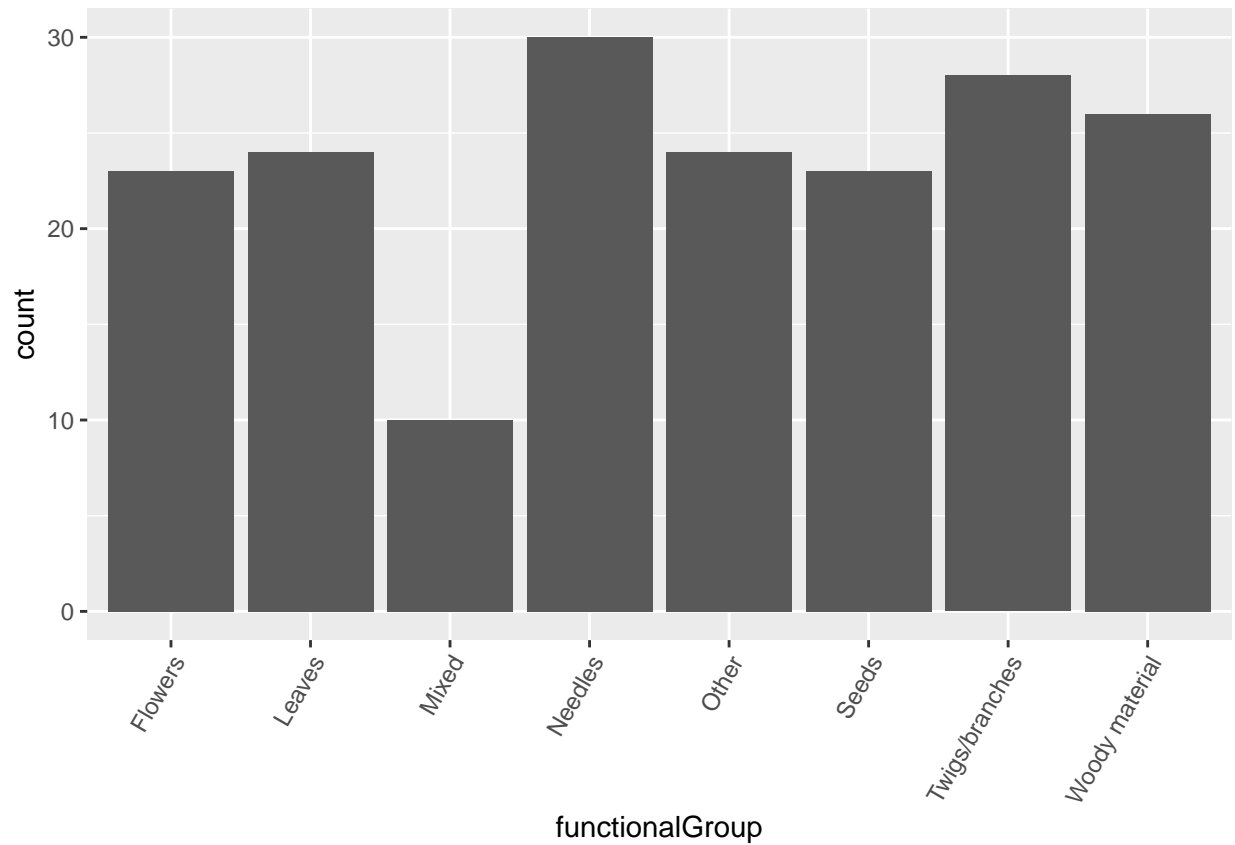
```
# compare with summary  
summary(Litter$siteID)
```

```
## NIWO  
## 188
```

Answer: From `unique` function, the results show all plots all sampled at Niwot Ridge (NIWO), so there are 188 plots in total. The difference of `unique` and `summary` function is the `unique` function can't produce count directly.

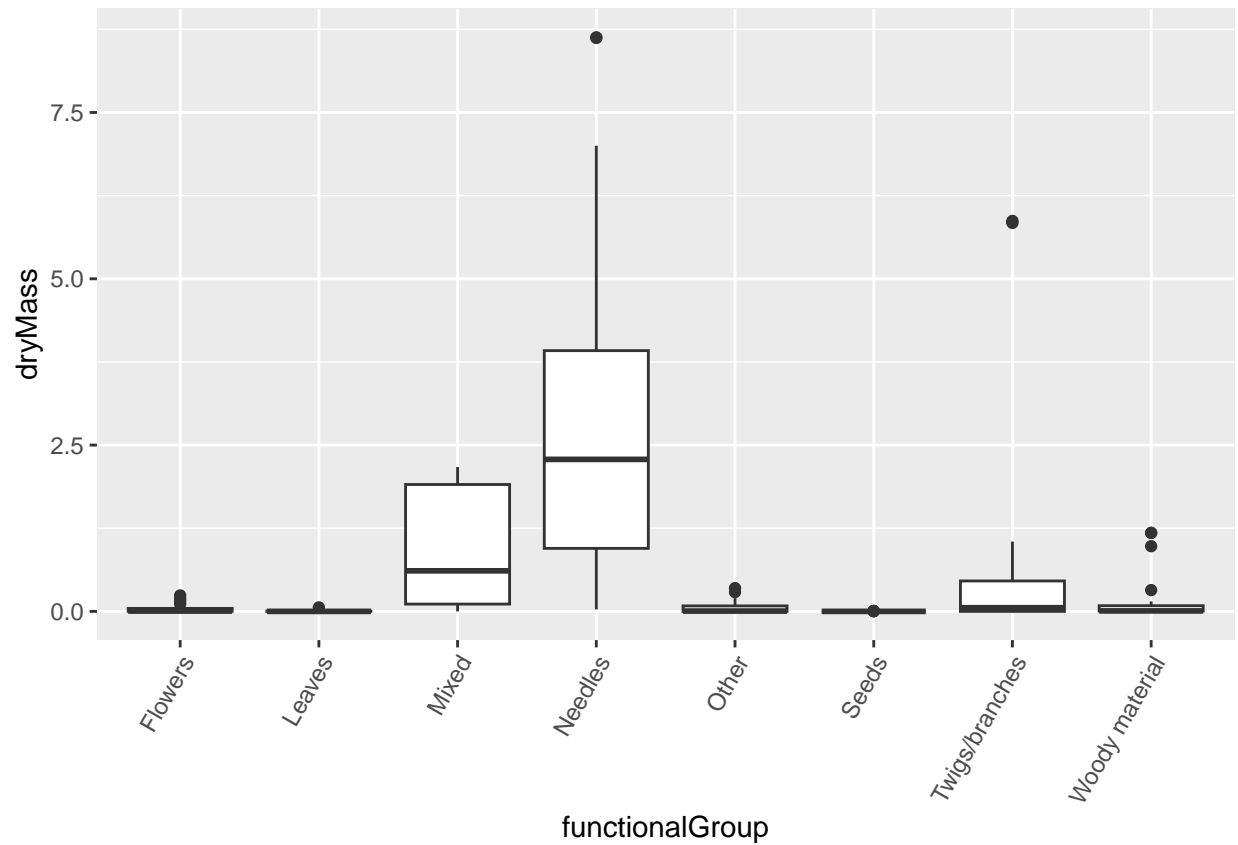
14. Create a bar graph of `functionalGroup` counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(x = functionalGroup)) +  
  geom_bar() + theme(axis.text.x = element_text(angle = 60, vjust = 1, hjust=1))
```

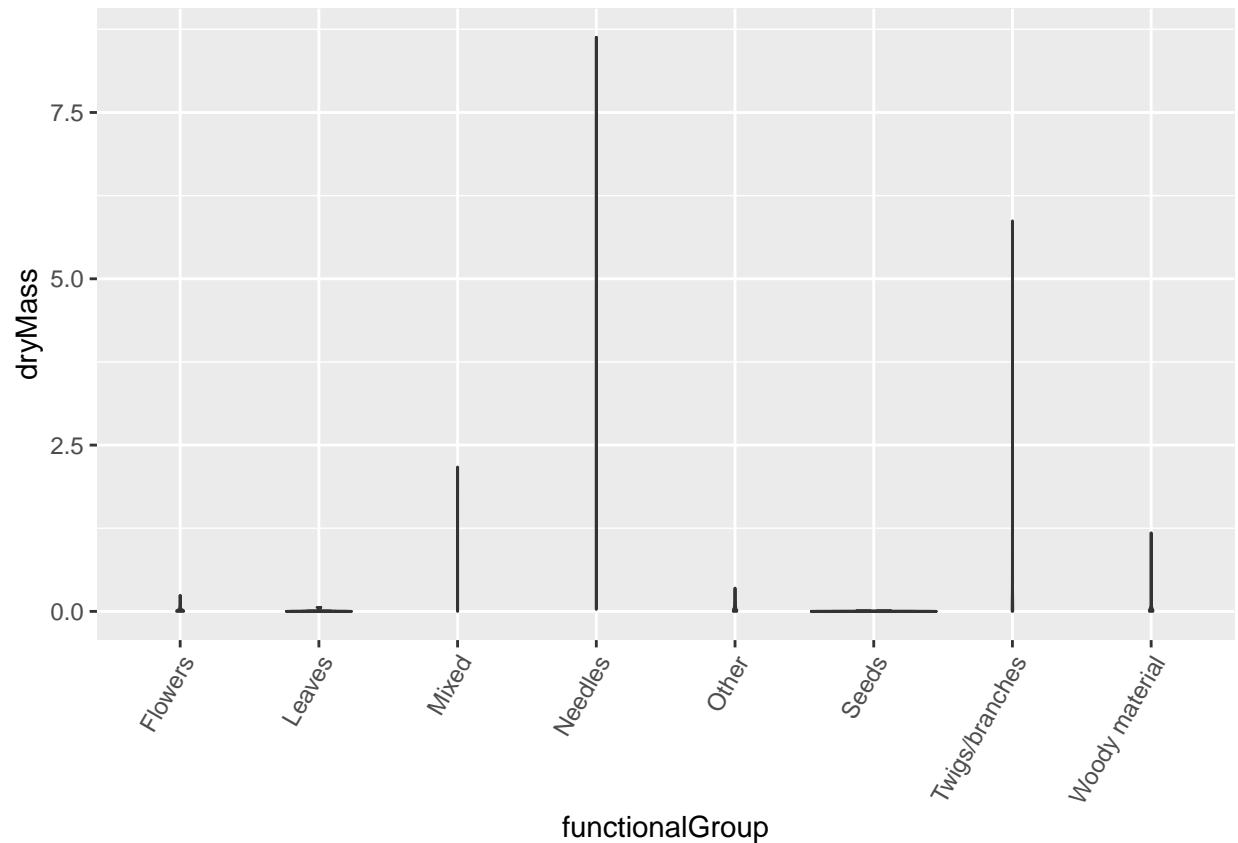


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of `dryMass` by `functionalGroup`.

```
ggplot(Litter) +  
  geom_boxplot(aes(x = functionalGroup, y = dryMass)) + # avoid axis label overlap  
  theme(axis.text.x = element_text(angle = 60, vjust = 1, hjust=1))
```



```
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
    draw_quantiles = c(0.25, 0.5, 0.75)) +
  theme(axis.text.x = element_text(angle = 60, vjust = 1, hjust=1))
```



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

Answer: There is low variability in the dry mass data. The dry mass value are almost unique so it will appear like a line in the frequency plot. So in this case, violin plot that can show frequency distribution is not as effective as boxplot that only shows distribution of values.

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

Answer: Needles. According to the boxplot, it has highest median and highest 25th to the 75th percentile value, also has the highest outlier value.