

Assignment 3: Data Exploration

Xuening Tang

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. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.
6. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai.

The completed exercise is due on Sept 30th.

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 include the subcommand to read strings in as factors.

```
getwd()
```

```
## [1] "/Users/xueningtang/Desktop/R lab/EDA-Fall2022/Assignments"
```

```
library(dplyr)
```

```
## Warning: replacing previous import 'lifecycle::last_warnings' by  
## 'rlang::last_warnings' when loading 'pillar'
```

```
library(ggplot2)  
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.2
```

```
## Warning: package 'tibble' was built under R version 4.1.2
```

```
## Warning: package 'tidyr' was built under R version 4.1.2
```

```
## Warning: package 'readr' was built under R version 4.1.2
```

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

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 are used on over 140 different agricultural crops in more than 120 countries. Due to the prevalence of using it, the danger to ecosystems and insects are not negelectable. They attack the central nervous system of insects, causing overstimulation of their nerve cells, paralysis and death, which could pose huge risk to the ecosystem balance and biodiversity.

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: Leaf litter and woody debris are important components of healthy soil. Decomposing litter and debris release nutrients into the soil and also keeps it moist. It also serves as great nesting material, hiding places and protected spots for animals. In this way, studying this dataset could help us develop more complete information about the forest 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: 1.Tower plots sampling equipment used as spatial sampling. 2.Place trap within plots either targeted or randomized based on the vegetation. 3.Sample ground traps every year as the temporal sampling.

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?

```
Neonics$Effect <- as.factor(Neonics$Effect)
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 effect studied is “Mortality”. Because the most severe effect for insects group is causing death and it could show how much toxic effects on insects.

- 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.

```
Neonics$Species.Common.Name <- as.factor(Neonics$Species.Common.Name)
class(Neonics$Species.Common.Name)
```

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

```
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: Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, Italian Honeybee are six most frequent species studied in this research. We can see they are all from species of bees, which is the most effected species by attacking the central nervous system of insects, causing overstimulation of their nerve cells, paralysis and death.

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.)
```

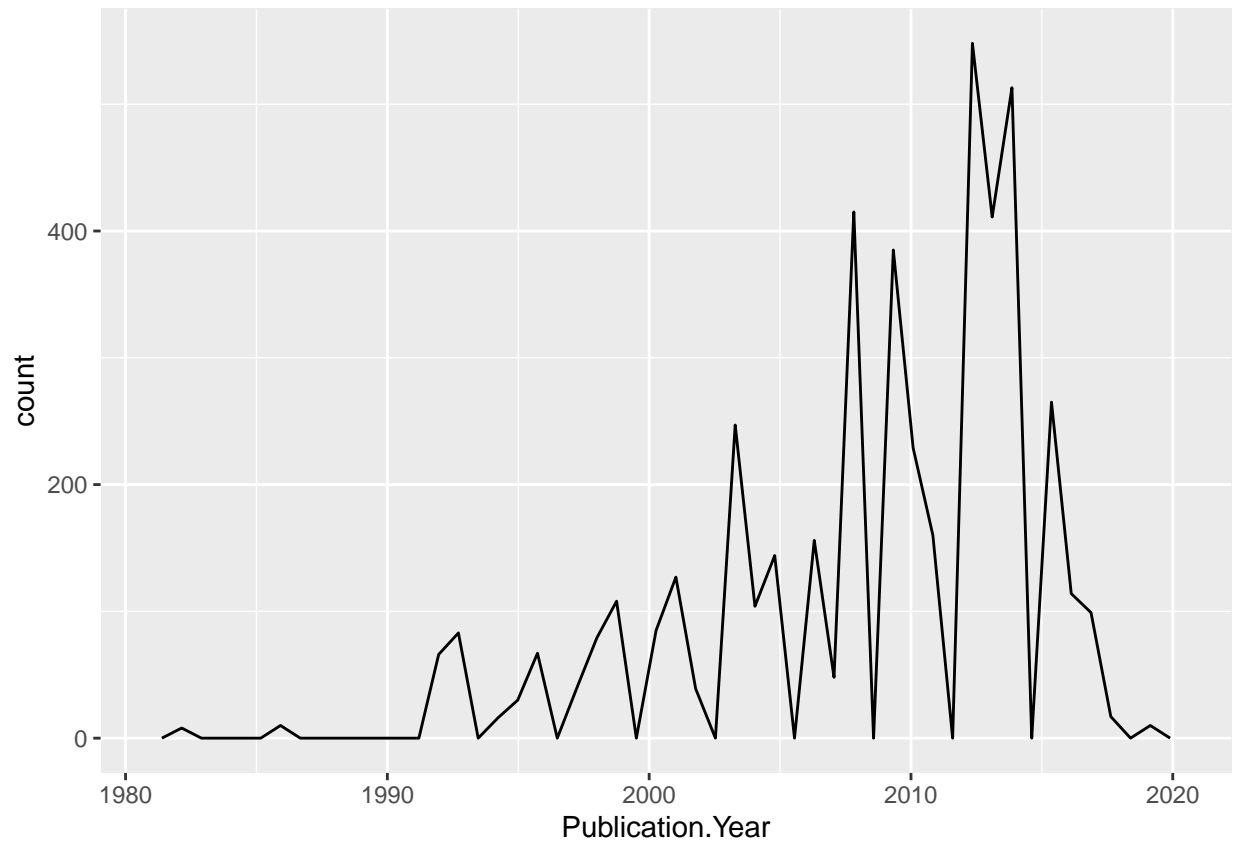
```
## [1] "character"
```

Answer: The class of Conc.1..Author is character, and it maybe because this column is separate from the previous column, which contains the concentraion and unit. Or it might due to its original format before putting in R has not be changed to number.

Explore your data graphically (Neonics)

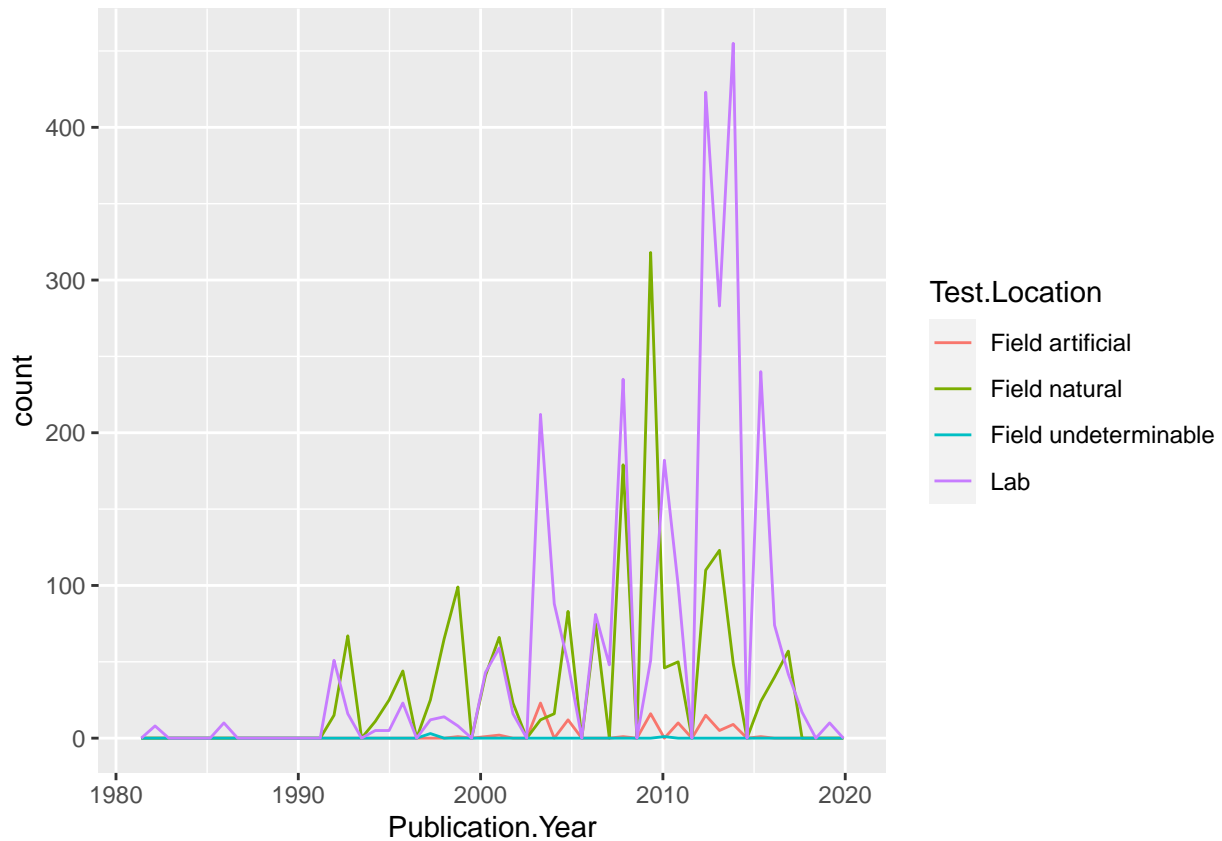
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year), bins = 50)
```



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, colour = Test.Location)) + geom_freqpoly(bins = 50)
```



```
scale_x_continuous(limits = c(1980, 2020))
```

```
## <ScaleContinuousPosition>
## Range:
## Limits: 1.98e+03 -- 2.02e+03
```

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

Answer: The most common test varies over time. At the first decade of collecting data, the main resource depends on the lab. Then in the next 100-year period, the most common test location becomes natural fields, which reached a peak at around 2009. After that, conducting experiments from lab gradually dominated back as the most common way to contain data until now.

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.

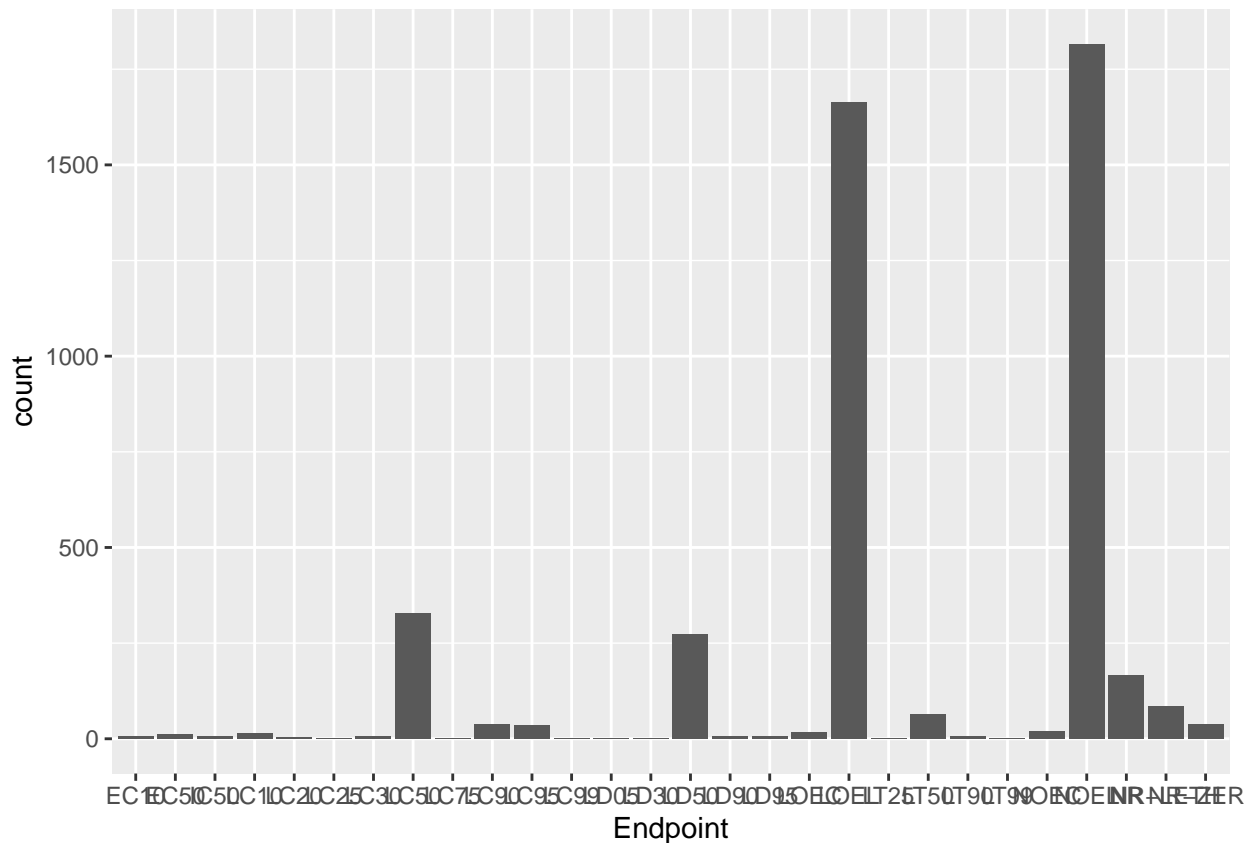
```
class(Neonics$Endpoint)
```

```
## [1] "character"
```

```
Neonics$Endpoint <- as.factor(Neonics$Endpoint)
summary(Neonics$Endpoint)
```

##	EC10	EC50	IC50	LC10	LC20	LC25	LC30	LC50	LC75	LC90
##	6	11	6	15	5	1	6	327	1	37
##	LC95	LC99	LD05	LD30	LD50	LD90	LD95	LOEC	LOEL	LT25
##	36	2	1	1	274	6	7	17	1664	1
##	LT50	LT90	LT99	NOEC	NOEL	NR	NR-LETH	NR-ZERO		
##	65	7	2	19	1816	167	86	37		

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



Answer: The two most common endpoints are population and mortality, which are defined by the effect. The population should be the insects population effected by the neonicotinoids and the mortality is defined as how many insects were infected to death compared as the whole population.

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] "character"
```



```
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"
## [7] "NIWO_047" "NIWO_051" "NIWO_058" "NIWO_046" "NIWO_062" "NIWO_057"
```

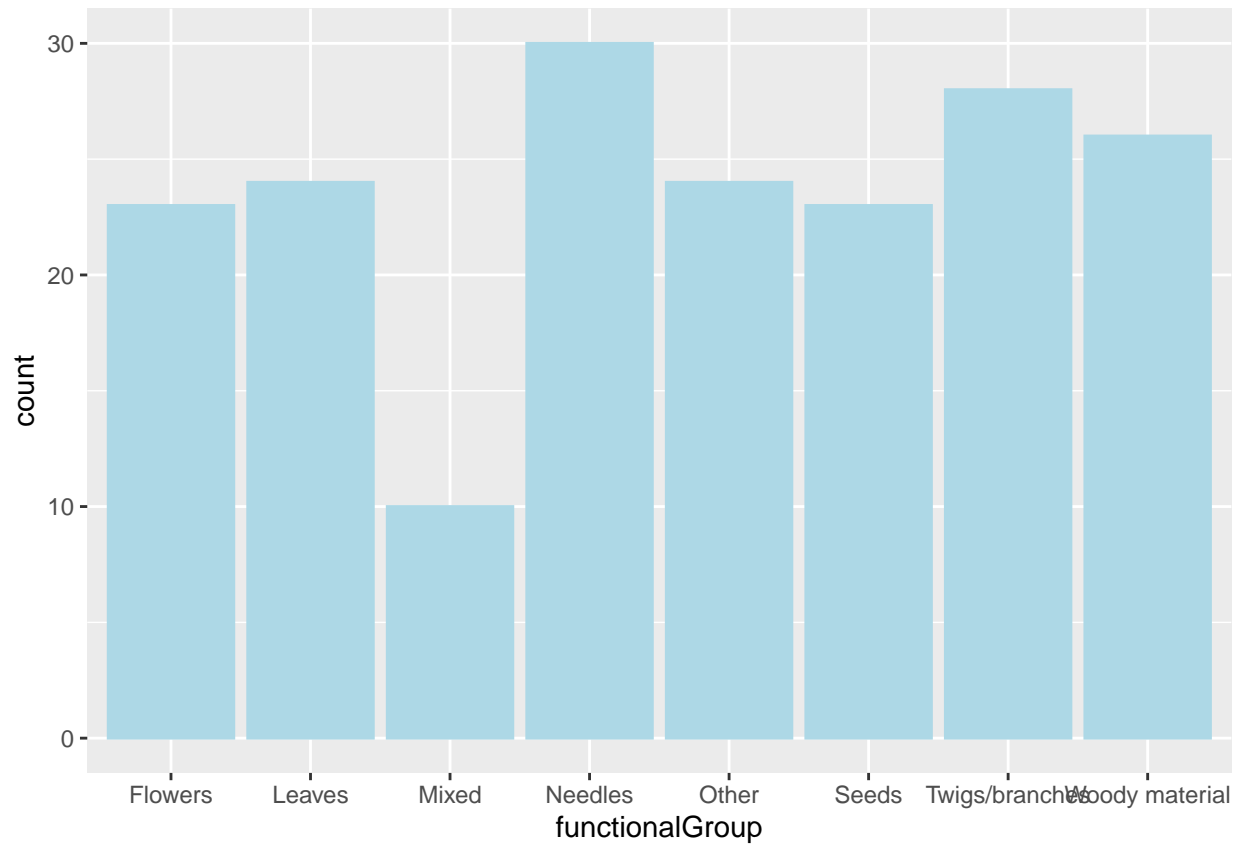
```
length(unique(Litter$plotID))
```

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

Answer: There are 12 plots sampled at Niwot Ridge. Data obtained from ‘`unique`’ could help to eliminate repeated values and help us understand the number of different character or value we have in this dataset. However, ‘`summary`’ function usually summarize the data to show its data distribution and characteristics by the class of the dataset.

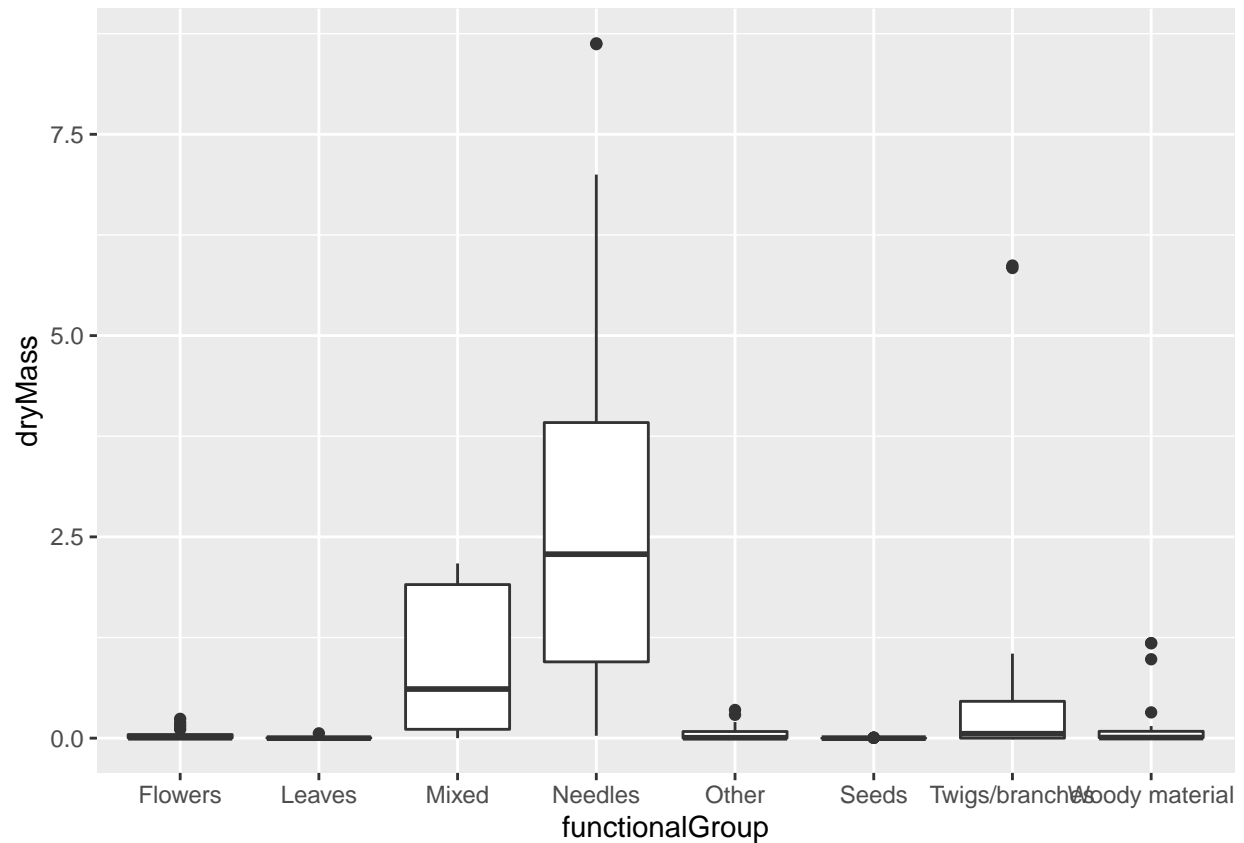
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(x = functionalGroup)) + geom_bar(color = "light blue", fill = "light blue")
```



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))
```

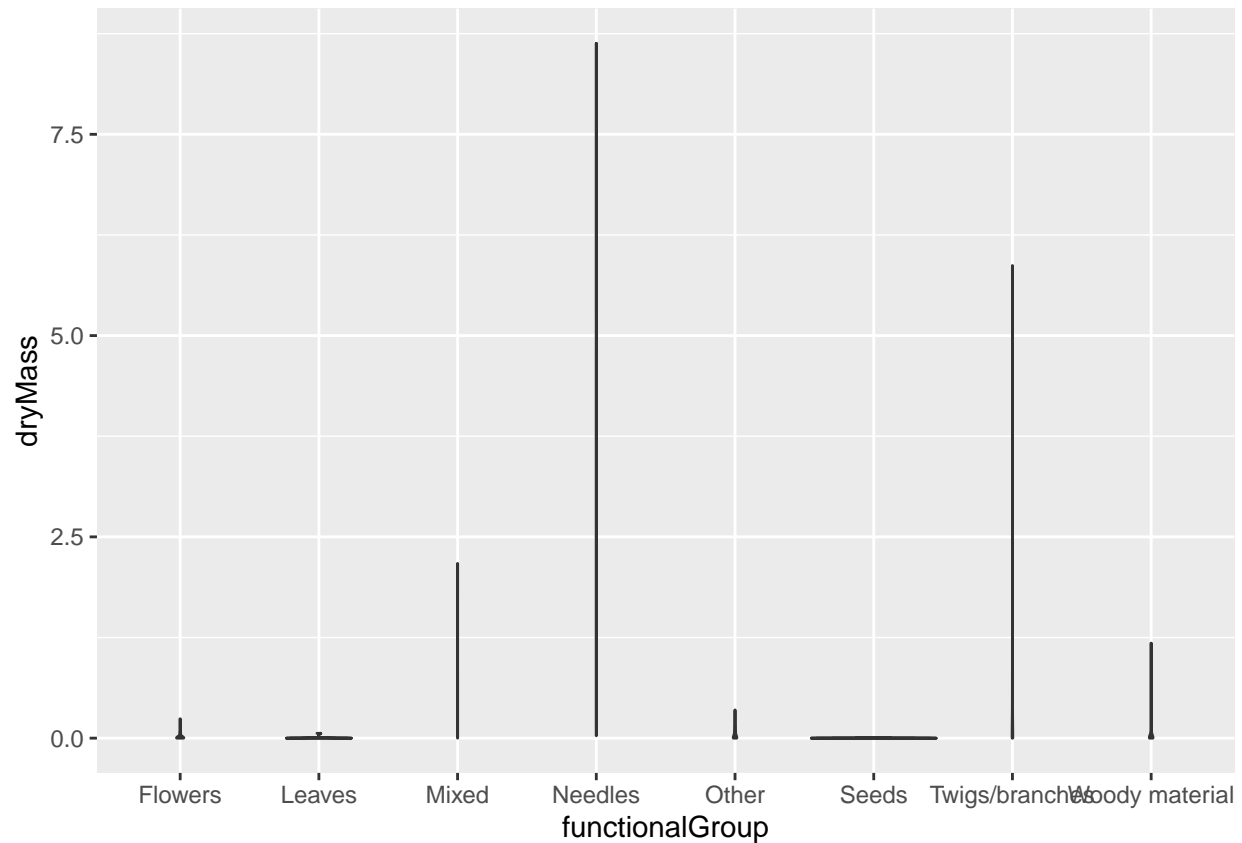


```
ggplot(Litter) + geom_violin(aes(x = functionalGroup, y = dryMass), draw_quantiles = c(0.25,
0.5, 0.75))
```

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```



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

Answer: In the box plot, it is clear to compare different groups' data distribution and understand the highest and the lowest biomass types. However, in violin plot, it can not show the direct distribution among one group from the low to high value because the function we used to eliminate repeated values by 'unique'.

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

Answer: The type of Needles have the highest median, mean and maximum value. Also, it distributed a lot more at the highest value zone than other types.