

Assignment 3: Data Exploration

Vanshika Mittal

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() #working directory is set
```

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

```
#install.packages("tidyverse")  
#loading necessary packages  
library(tidyverse)  
library(lubridate)
```

```
#uploading datasets
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",
  stringsAsFactors = TRUE, header = TRUE)
Litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv",
  stringsAsFactors = TRUE, header = 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: There has been a large evidence of the harm caused from chemicals such as neonicotinoids present usually in insecticides. This includes harm caused to insects.

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: Natural wood debris and forest litter is integral for regeneration of soil quality and reforestation.

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. Data from different collection events are measured separately. 2. That material is collected in 0.5m² PVC traps which is consistent with the the Smithsonian Tropical Research Institute Center for Tropical Forest Science 3. The location of the plots are selected randomly.

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: Population is the most studied effect. This could be the case due to rapidly decreasing and several insect species going extinct in recent times.

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...]

```
#summarising and sorting in descending order to find the most commonly studied species
summary(Neonics$Species.Common.Name, decreasing = FALSE)
```

##	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 most commonly studied insects are Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee and Italian Honeybee. They all belong to the family of bees and are known to be the most efficient pollinators. Bees are also endangered and are heavily impacted by insecticides.

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.) #checking the class of a column
```

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

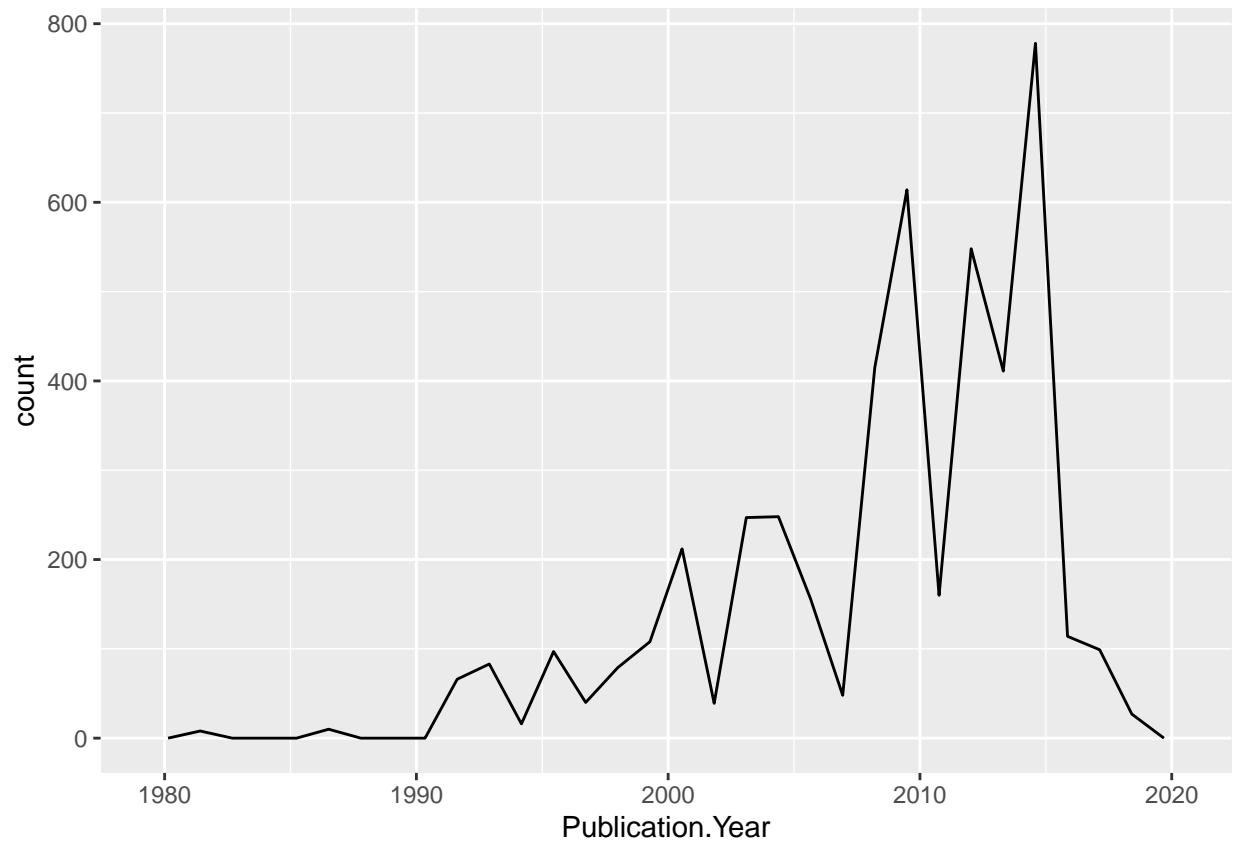
Answer: There are certain cells in this column which contain the value “NR” which could be the possible reason behind R reading it as a string value. While loading the data into our script, we changed as string values to be read as factors

Explore your data graphically (Neonics)

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

```
#graphing the frequency of publications by year
Neonics_Studies_Frq1 <- ggplot(Neonics) +
  geom_freqpoly(aes(x=Publication.Year))
Neonics_Studies_Frq1 # calling the frequency polygraph
```

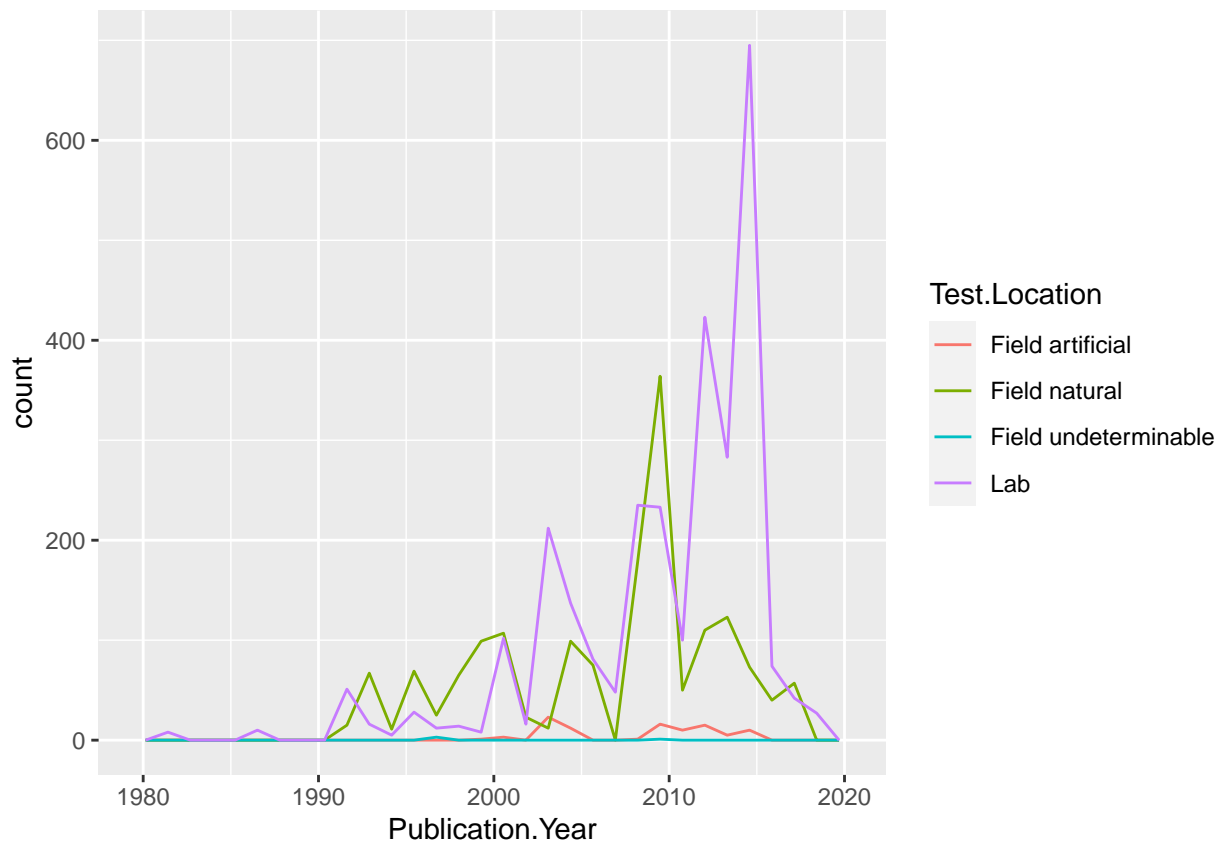
```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



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

```
#graphing the frequency of publications by year and sorting them in different colours for different tes
Neonics_Studies_Frq2 <- ggplot(Neonics) +
  geom_freqpoly(aes(x=Publication.Year, color = Test.Location))
Neonics_Studies_Frq2 #calling the frequency polygraph
```

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



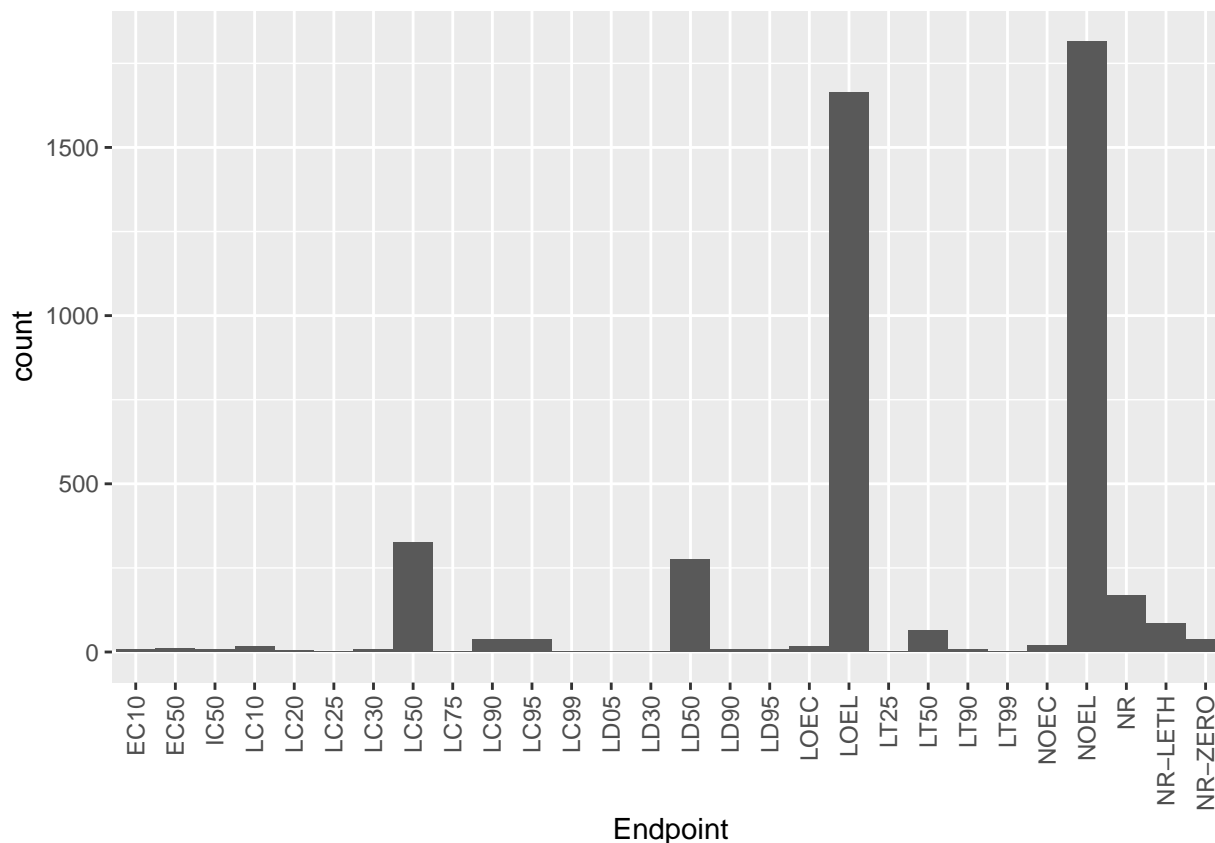
Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: The most common test locations are Lab and then Field natural. While field natural peaked before 2010, lab sites have picked up in frequency more recently and gained extreme popularity in the last decade.

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...]

```
#making a bar graph
Neonics_Endpoint_Bar <- ggplot(Neonics, aes (x=Endpoint, after_stat(count))) +
  geom_bar(aes(x=Endpoint, after_stat(count)), width = 1)+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
Neonics_Endpoint_Bar #calling the bar graph
```



Answer: Most common end points are LOEL and NOEL. LOEL stands for Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC). NOEL stands for No-observable-effect-level: highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test (NOEAL/NOEC) *Definitions taken from the the ECOTOX_CodeAppendix

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) #determining class of a column
```

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

```
summary(Litter$collectDate) # looking at the way the date is defined
```

```
## 2018-08-02 2018-08-30
##          91          97
```



```
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d") # formatting the date to change class
class(Litter$collectDate) #confirming the change in variable class
```

```
## [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$siteID)
```

```
## [1] NIWO
```

```
## Levels: NIWO
```

```
summary(Litter$siteID)
```

```
## NIWO
```

```
## 188
```

Answer: The `unique` function does not report the number of plots and only the unique values that the variable takes. On the other hand, `summary` reports the frequency of the different values.

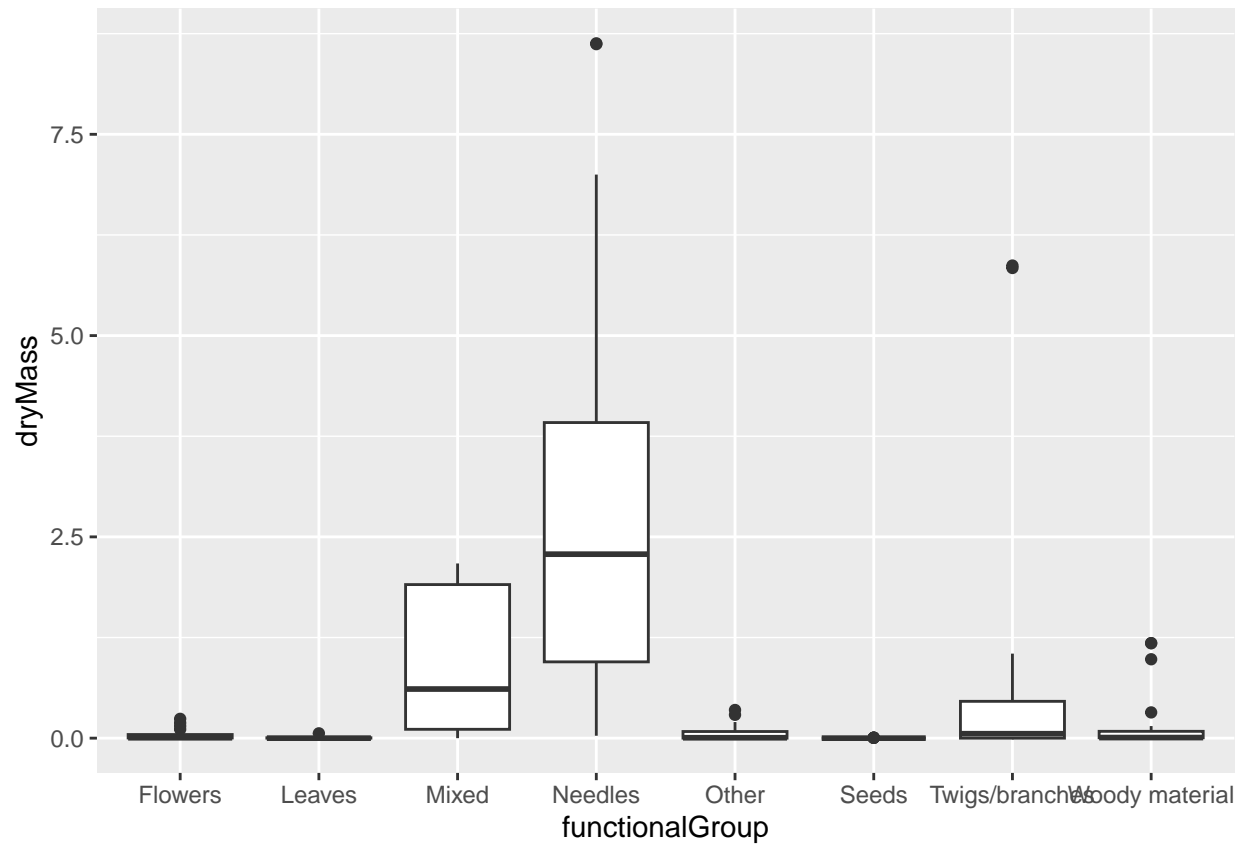
14. Create a bar graph of `functionalGroup` counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#plotting a bar graph for the type of litter collected
Litter_functional_Bar <- ggplot(Litter)+
  geom_bar(aes(x=functionalGroup, after_stat(count)))
Litter_functional_Bar
```

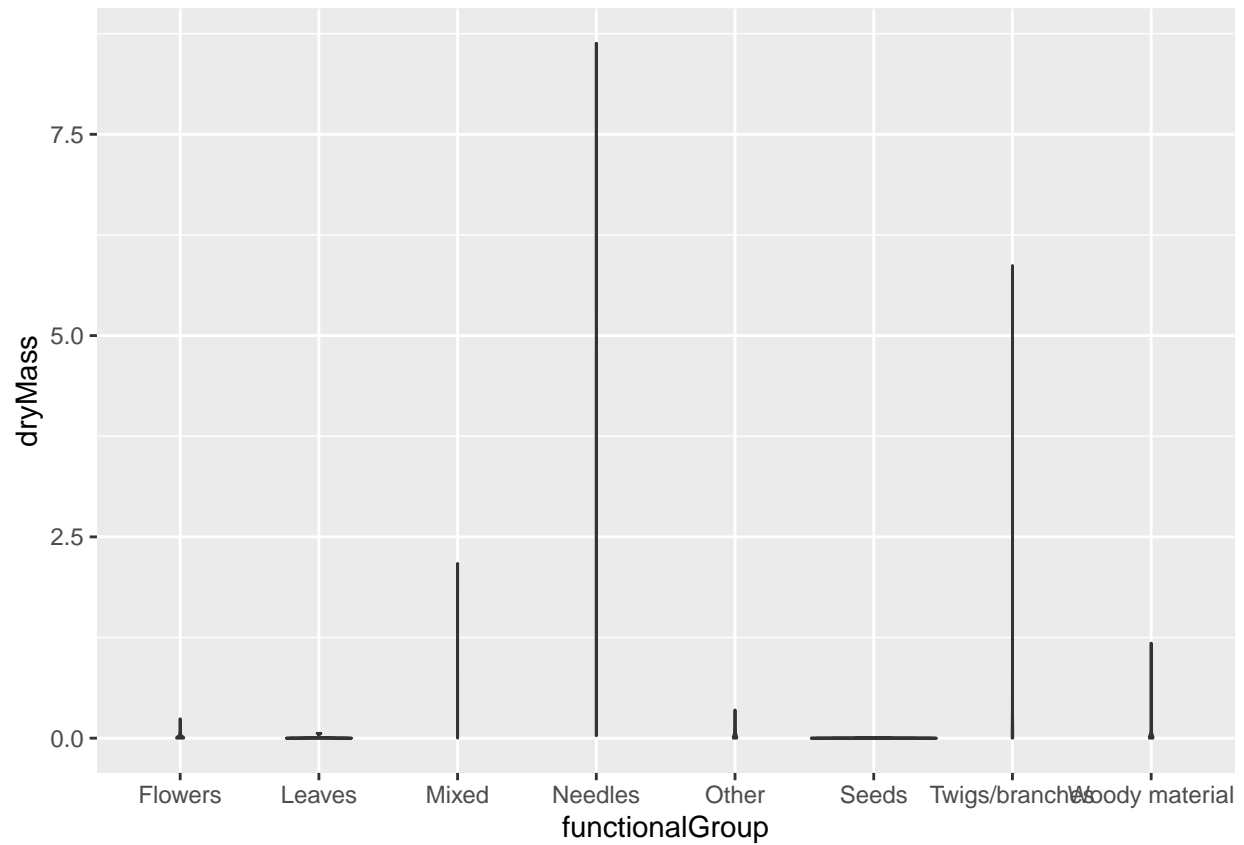


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of `dryMass` by `functionalGroup`.

```
#making a boxplot
Litter_functional_Boxplot <- ggplot(Litter)+
  geom_boxplot(aes(x = functionalGroup, y = dryMass))
Litter_functional_Boxplot
```



```
#using the violinplot to visualise the data
Litter_functional_Violin <- ggplot(Litter)+
  geom_violin(aes(x = functionalGroup, y = dryMass),
    draw_quantiles = c(0.25, 0.5, 0.75))
Litter_functional_Violin
```



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

Answer: In this case, we are using qualitative data which makes the boxplot a more suitable option to view the distribution of our data.

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

Answer: Litter classified under Needles has the highest biomass at these sites.