

Assignment 3: Data Exploration

Nicole Eastman

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.

```
setwd("/home/guest/R/EDA Fall 2022")
getwd()
```

```
## [1] "/home/guest/R/EDA Fall 2022"
```

```
library(tidyverse)
ECOTOX.insects.data <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = TRUE)
NEON.litter.debris.data <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv", stringsAsFactors = TRUE)
Neonics <- ECOTOX.insects.data
Litter <- NEON.litter.debris.data
```

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: Scientists may be interested in this dataset representing the ecotoxicology of neonicotinoids on insects because it will assist in evaluating which substance is effective in killing insects that may be harming agricultural products. This information can inform farmers on which neonicotinoids to use on their farms.

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: We would be interested in studying litter and woody debris on the forest floor because it serves as habitat for insects living in that ecosystem. Therefore, we can assess available habitat and food sources for insects. This organic matter also contributes to the nutrient cycle and carbon budgets.

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. The sampling of litter and fine woody debris occurs in tower plots. The sites with forested tower airsheds sample in 20 40mx40m plots while the sites with low-statured vegetation are sampled in 4 40mx40m and 26 20mx20m plots. 2. Each plot has 1-4 trap pairs with one elevated trap and one ground trap in each pair. They are deployed in every 400m² of plot area. 3. One time each year, ground traps are sampled; contrastingly, elevated traps are sampled once bi-weekly in deciduous forest sites during senescence and once every 1 to 2 months year-long in coniferous forest sites.

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: The most common effect is population with a value of 1803. This may be the result of the experiment being conducted on a group of insects of the same species at the same time and place. This effect is of interest because it is important to know what type of insect species had a given reaction to the chemical application. Additionally, mortality was a frequent effect which had a count of 1493, meaning the application of the chemical resulted in death of the insect. This is an important effect to evaluate because the goal of the study is to assess which chemicals are effective in killing insects.

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 include Honey Bees, Parasitic Wasps, Buff Tailed Honeybees, Carniolan Honey Bees, Bumble Bees, and Italian Honey Bees. All of these species are types of bees which may be of interest because they tend to be species you want for pollination purposes. If these species die in response to the insecticide, the farmer may want to chose a different chemical.

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.)
```

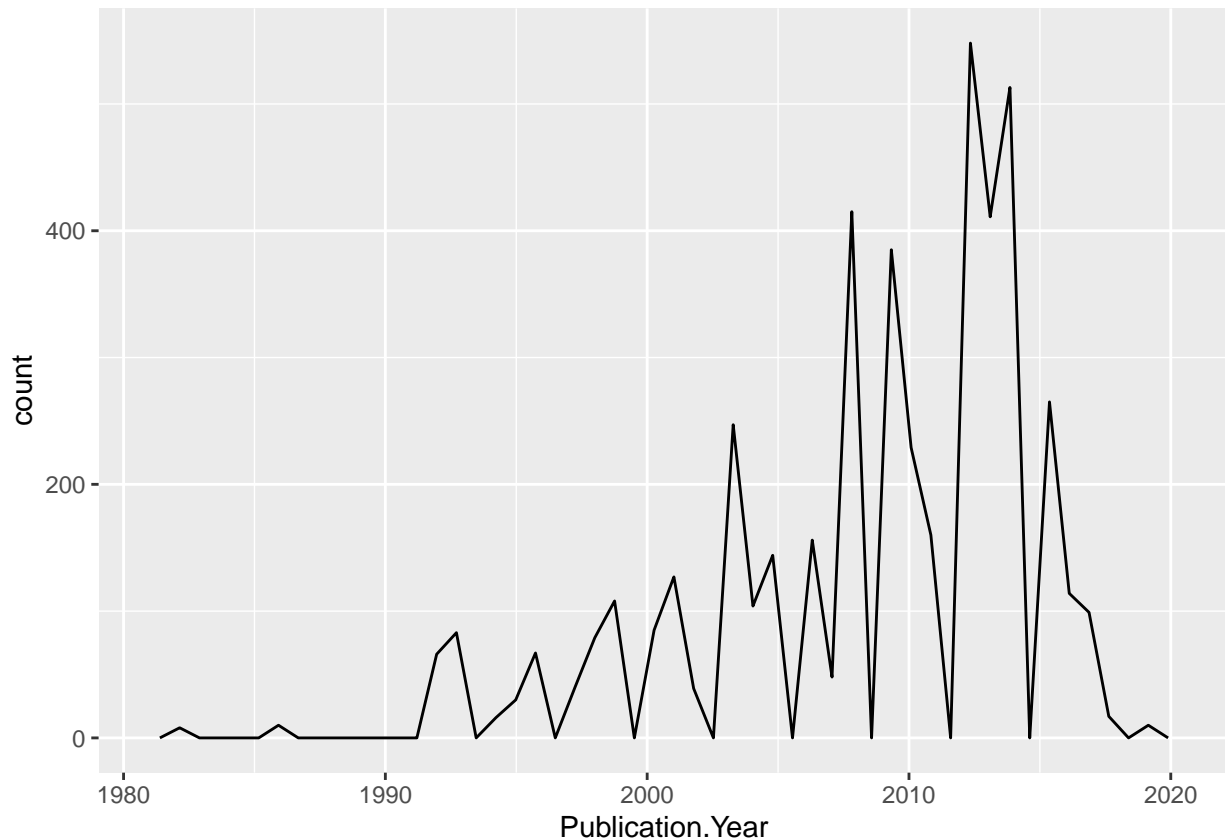
```
## [1] "factor"
```

Answer: The class of `Conc.1..Author.` is a factor because it represents levels or categories of concentrations rather than individual numeric values for each row.

Explore your data graphically (Neonics)

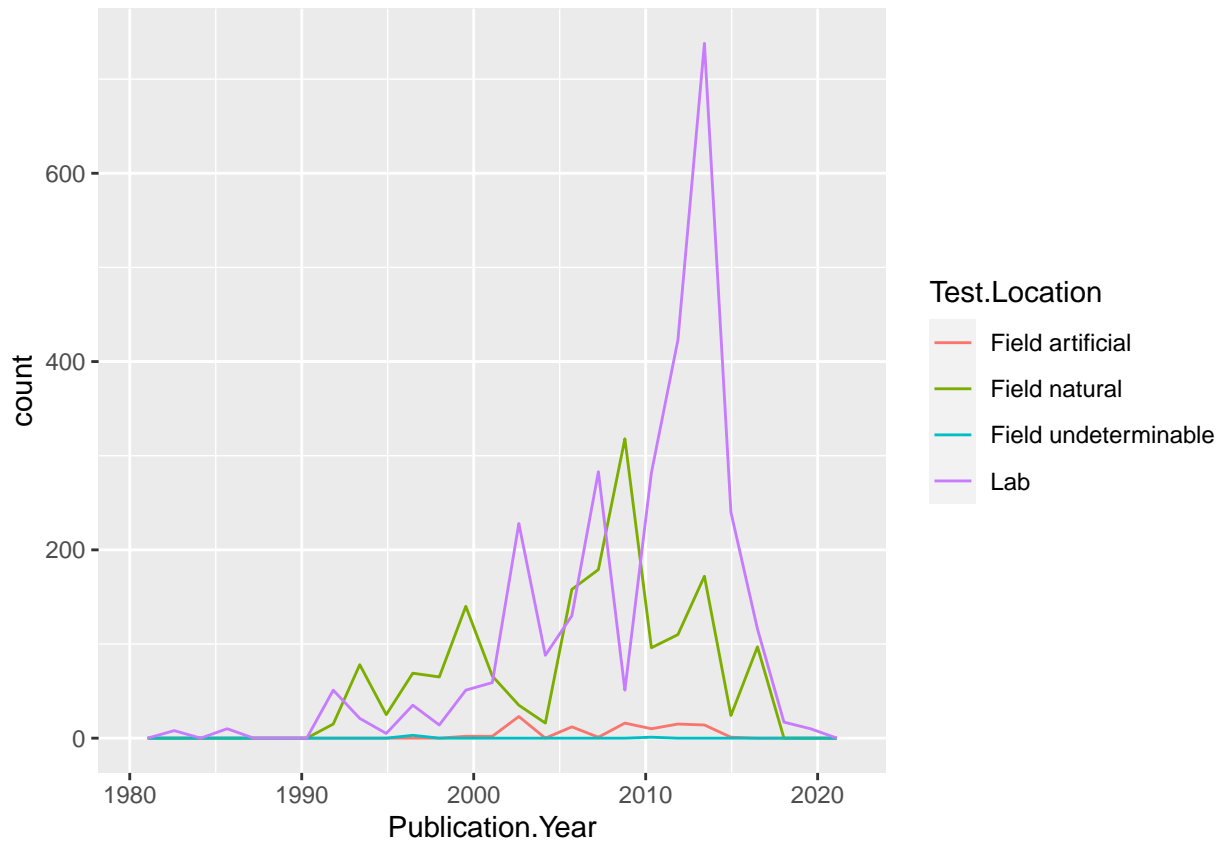
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
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) + geom_freqpoly(aes(x = Publication.Year, color = Test.Location),  
  bins = 25)
```

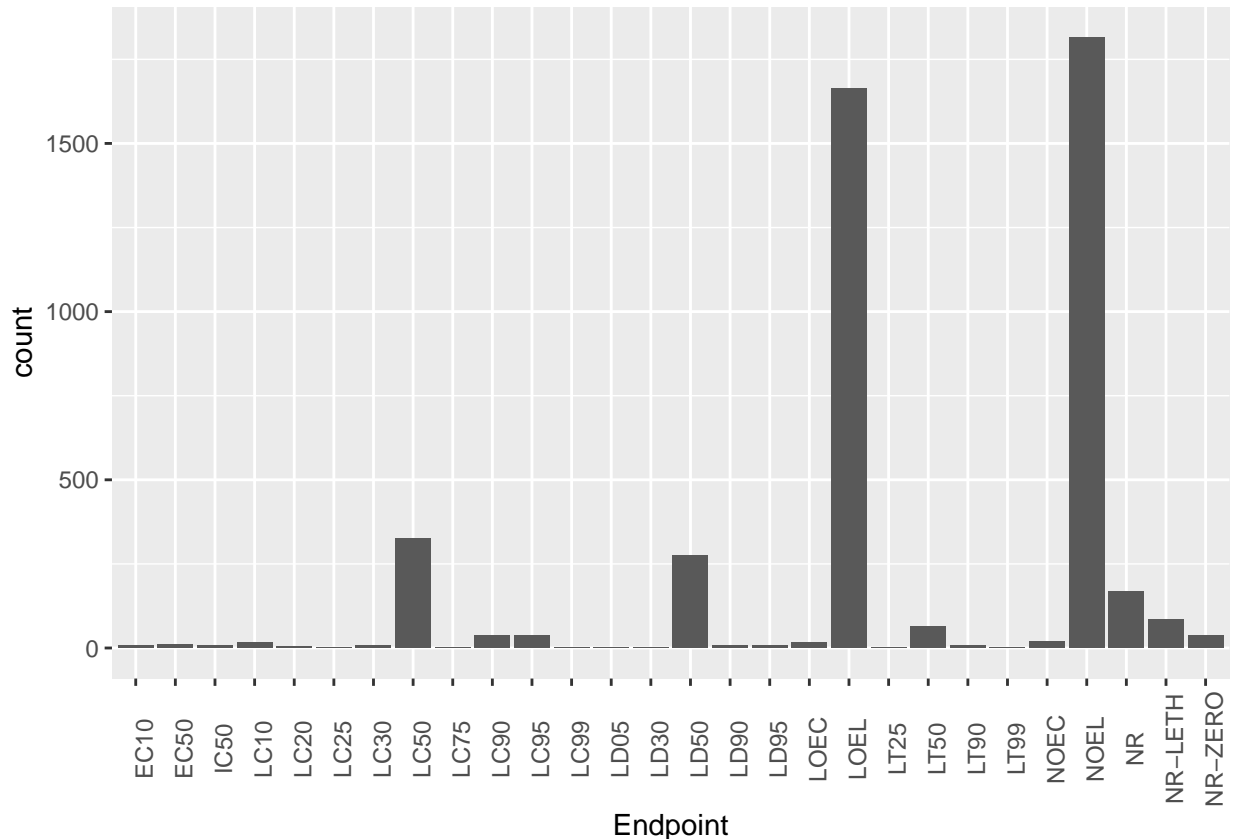


Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: The most common test locations are in labs and the natural field. The testing frequency begins to increase in the 1990s and peaks around 2013-2014.

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() + theme(axis.text.x = element_text(angle = 90))
```



Answer: The most common endpoint is NOEL meaning the no-observable-effect-level. This is defined as the highest concentration that produces an effect that is not significantly different from the responses of the control. LOEL has the second greatest count and is defined as the lowest-observable-effect-level which means the lowest concentration that produces effects that are different from the lowest-observable-effect concentration control.

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")
```

```
unique(Litter$collectDate)
```

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

- 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$namedLocation)
```

```
## [1] NIWO_061.basePlot.ltr NIWO_064.basePlot.ltr NIWO_067.basePlot.ltr
```

```
## [4] NIWO_040.basePlot.ltr NIWO_041.basePlot.ltr NIWO_063.basePlot.ltr
```

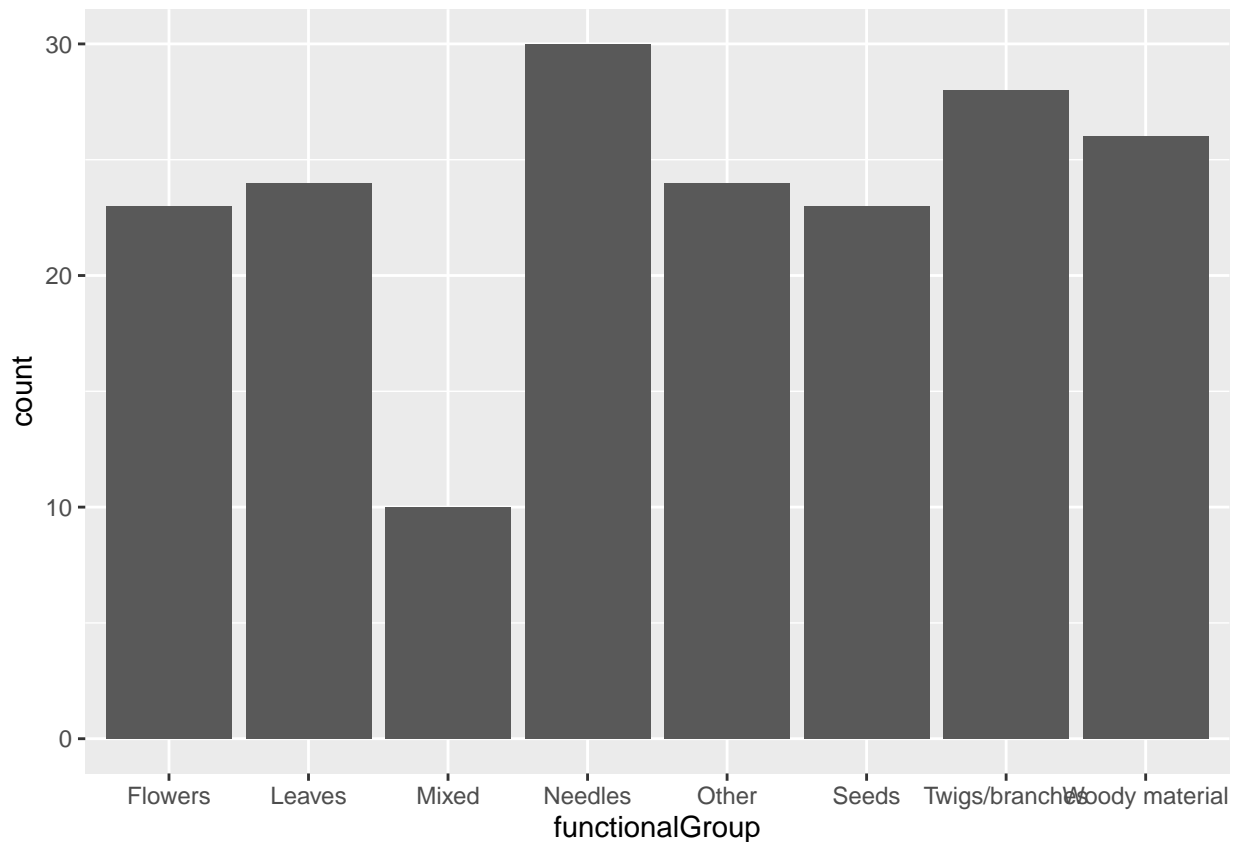
```
## [7] NIWO_047.basePlot.ltr NIWO_051.basePlot.ltr NIWO_058.basePlot.ltr
```

```
## [10] NIWO_046.basePlot.ltr NIWO_062.basePlot.ltr NIWO_057.basePlot.ltr
## 12 Levels: NIWO_040.basePlot.ltr ... NIWO_067.basePlot.ltr
```

Answer: 12 different plots were sampled at Niwot Ridge. 'Unique' differs from 'summary' because it identifies each unique value in a row and categorizes it by the amount of duplicates for each value. Meanwhile, summary lists every value present in a row and the count for each value.

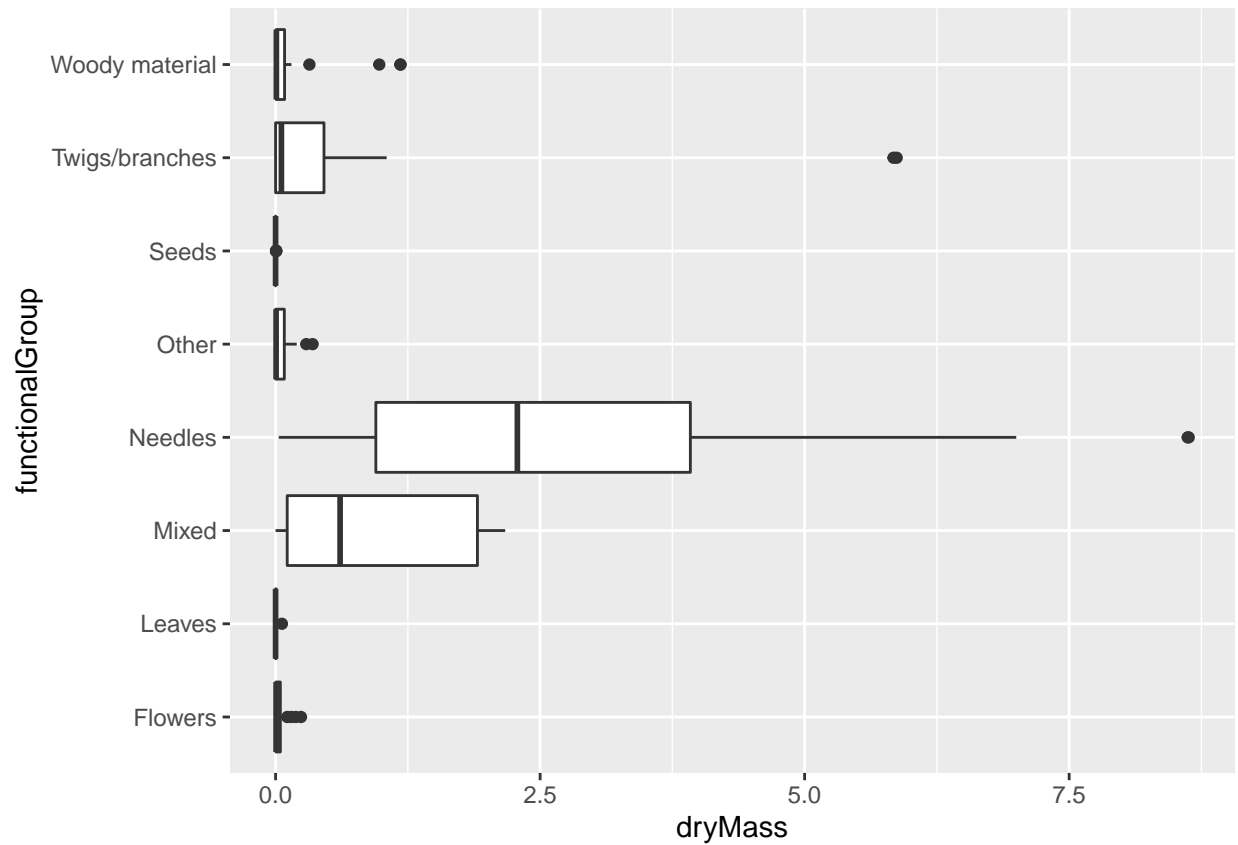
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
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 = dryMass, y = functionalGroup))
```

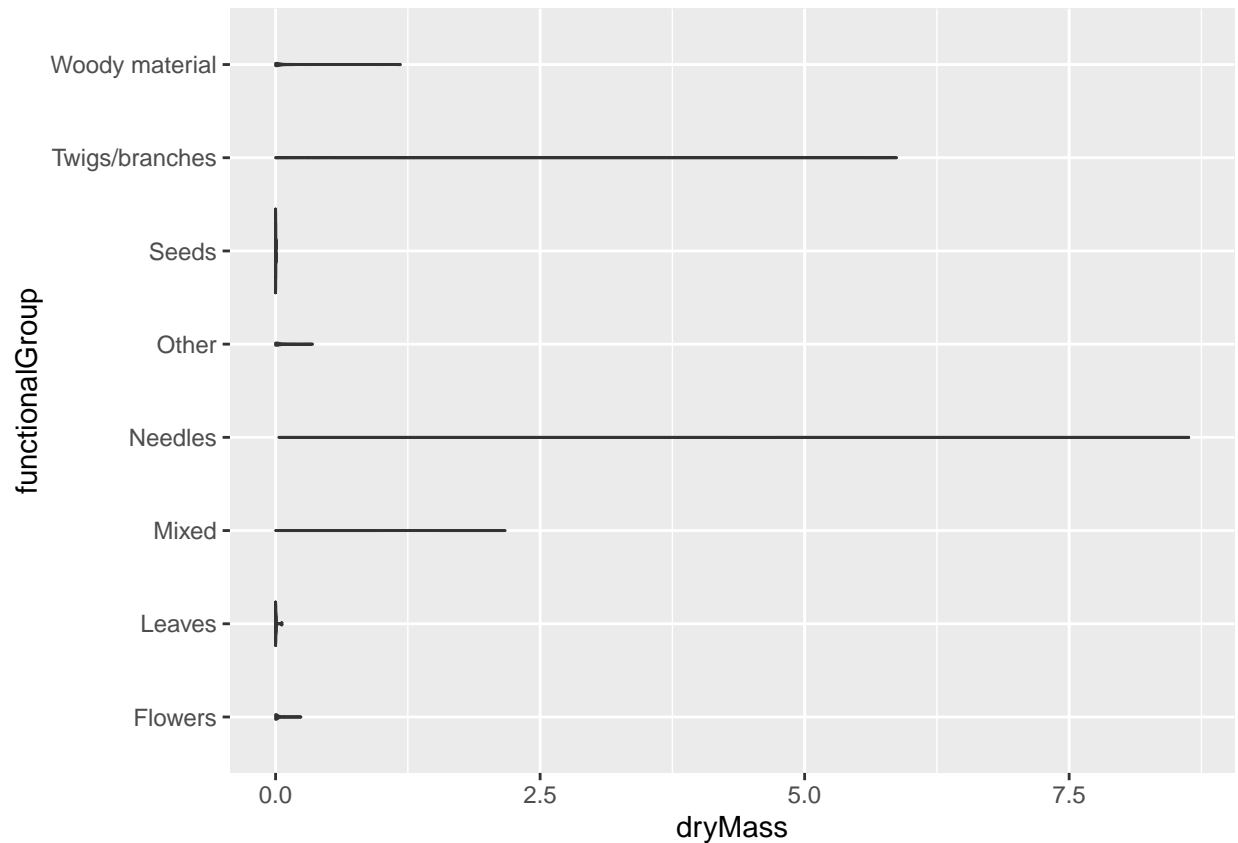



```
ggplot(Litter) + geom_violin(aes(x = dryMass, y = functionalGroup), draw_quantiles = c(0.25,
0.5, 0.75))
```

```
## 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: The boxplot is a more effective visualization because it shows the frequency of dryMass in addition to just the range of values. The median and quartiles are shown which indicate where 75% of the data for each functionalGroup occurs.

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

Answer: Needles tend to have the highest biomass at this site with the greatest median, range, and outlier.