

Assignment 3: Data Exploration

Julia Weinberg, Section #0

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.**

```
getwd()

## [1] "/Users/juliaweinberg/Desktop/github repos/Environmental_Data_Analytics_2022/Assignments"

#install.packages(tidyverse)
library(tidyverse)
#install.packages(lubridate)
library(lubridate)

Neonics <- read.csv("../Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = TRUE) #upload
Litter <- read.csv("../Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv", stringsAsFactors = TRUE) #upload
```

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 have been shown to have adverse impacts on pollinators. A reduction in pollinator species would lead to food scarcity on the planet, so assessing the ecotoxicology of neonicotinoids would be beneficial to protect insect species, not only limited to pollinators.

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 might be interested in studying leaf litter and woody debris because it allows us to see the organic content that will decompose into the top level of the soil. If the leaf litter is high in nutrients, then the soil will be nutrient dense and be able to support a thriving 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: *Ground traps are sampled yearly.* Elevated traps are sampled biweekly in deciduous forests and every 1-2 months in evergreen sites. *Elevated traps in deciduous locations are often not sampled for up to six months during the dormant winter seasons.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) #get dimensions of 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) #get summary of Neonics Effect column
```

```
##      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 commonly studied effects are mortality, behavior, feeding behavior, and population. These are of specific interest because they give insight into the survival rate of insect species as a result of the pesticide. Since species extinction is something that might be of concern, these are especially important parameters to look at.

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) #summary of Neonics species
```

```
##              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 species include: Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. All of these species are polinators and they are of increased interest because they are necessary for plant cultivation.

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.) #determining class of Conc.1..Author
```

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

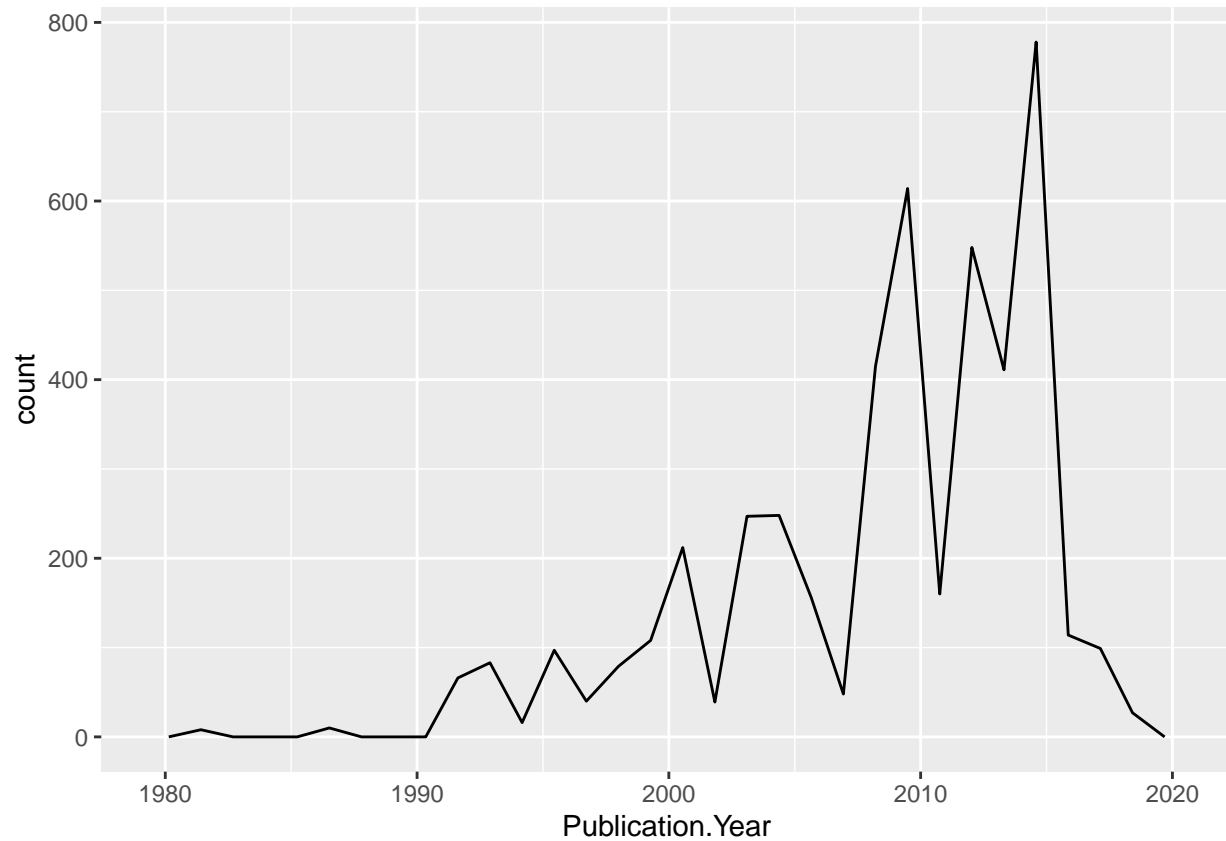
Answer: Conc.1..Author is factor data. It is not numeric because it is representing categorical data.

Explore your data graphically (Neonics)

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

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

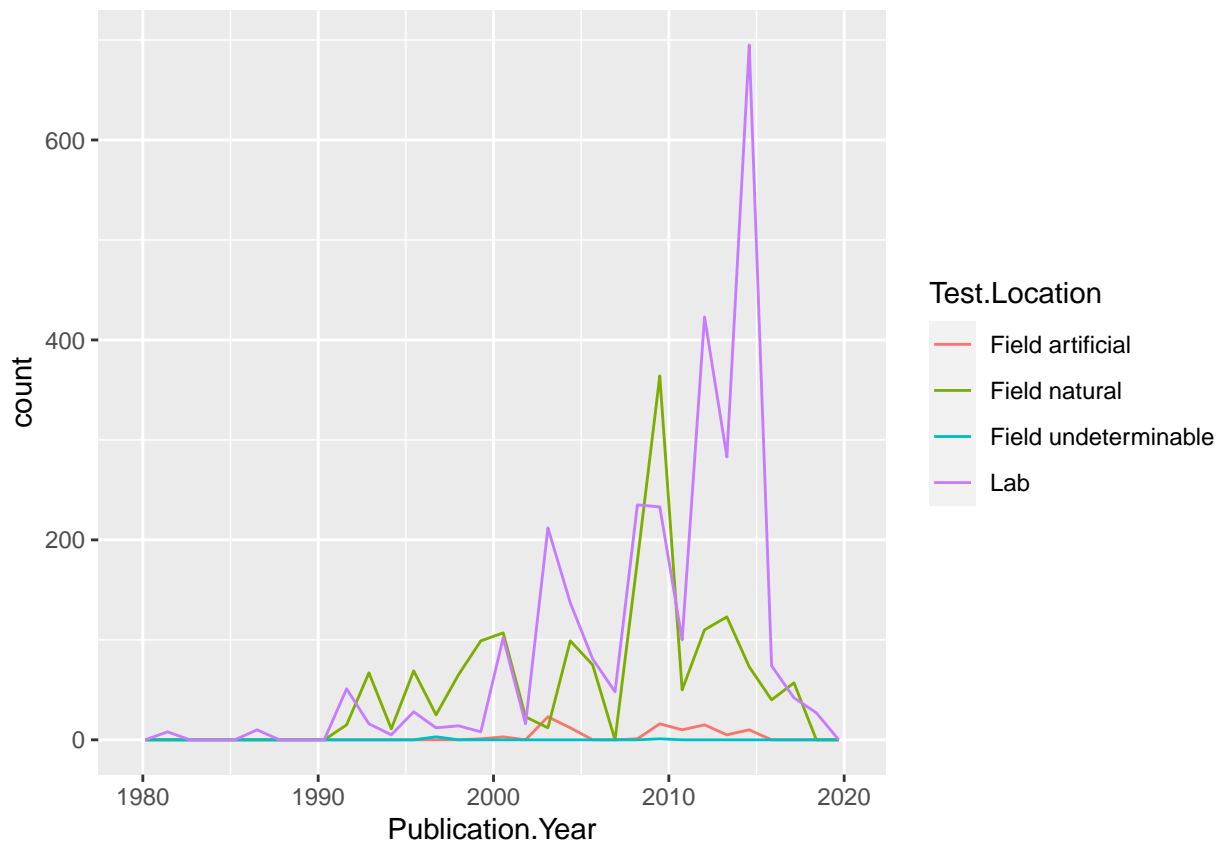
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
ggplot(Neonics, aes(x = Publication.Year, color = Test.Location)) + geom_freqpoly()
```

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

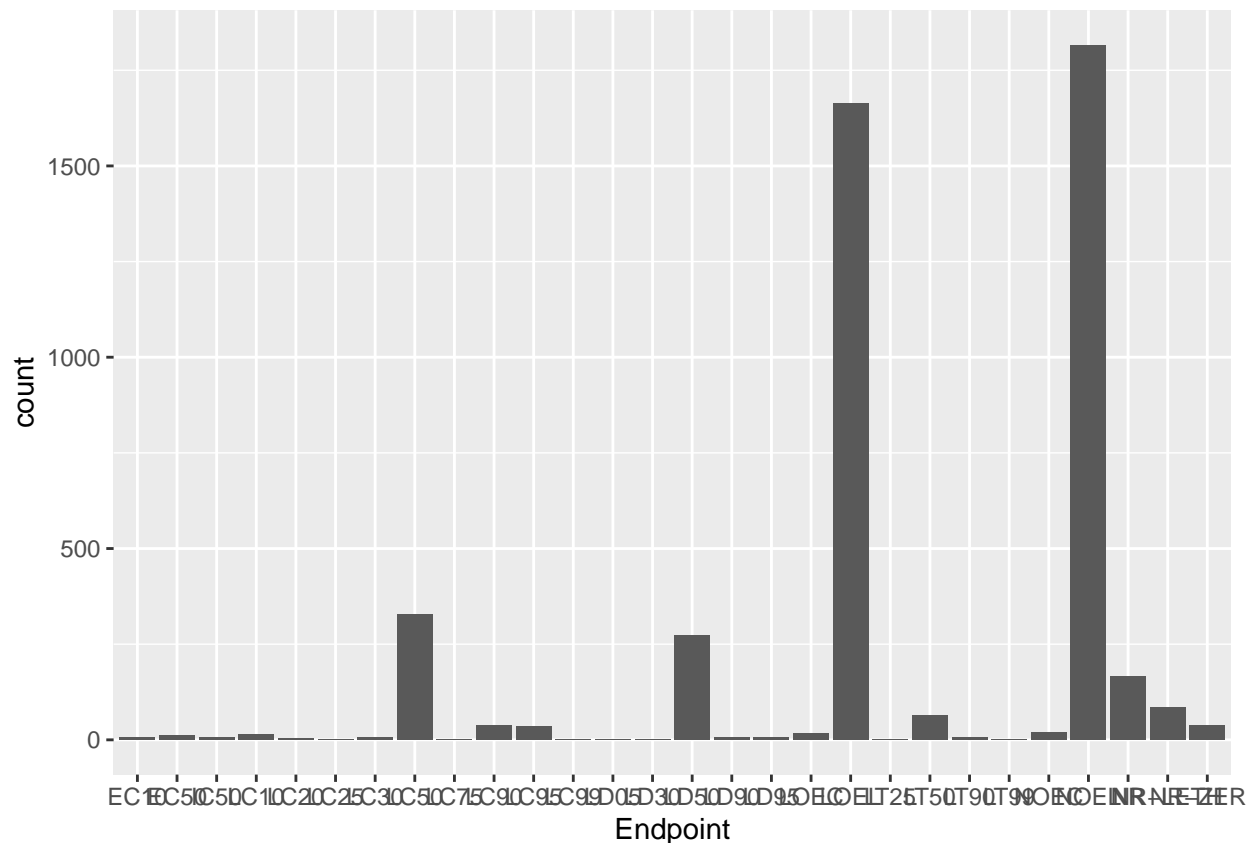


Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: The most common

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() #create bar graph of endpoint frequency
```



```
summary(Neonics$Endpoint) #improve readability of axes
```

```
##      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
```

Answer: The two most common end points are LOEL and NOEL. These are defined as

Explore your data (Litter)

12. 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) #determine class of collectDate
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate, format = "%M/%D/%Y")
#change collectDate to date class
```

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

```
## [1] NA
```

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`?

```
df_unique <- unique(Litter$plotID)
length(df_unique) #determine how many plots sampled at Niwot Ridge
```

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

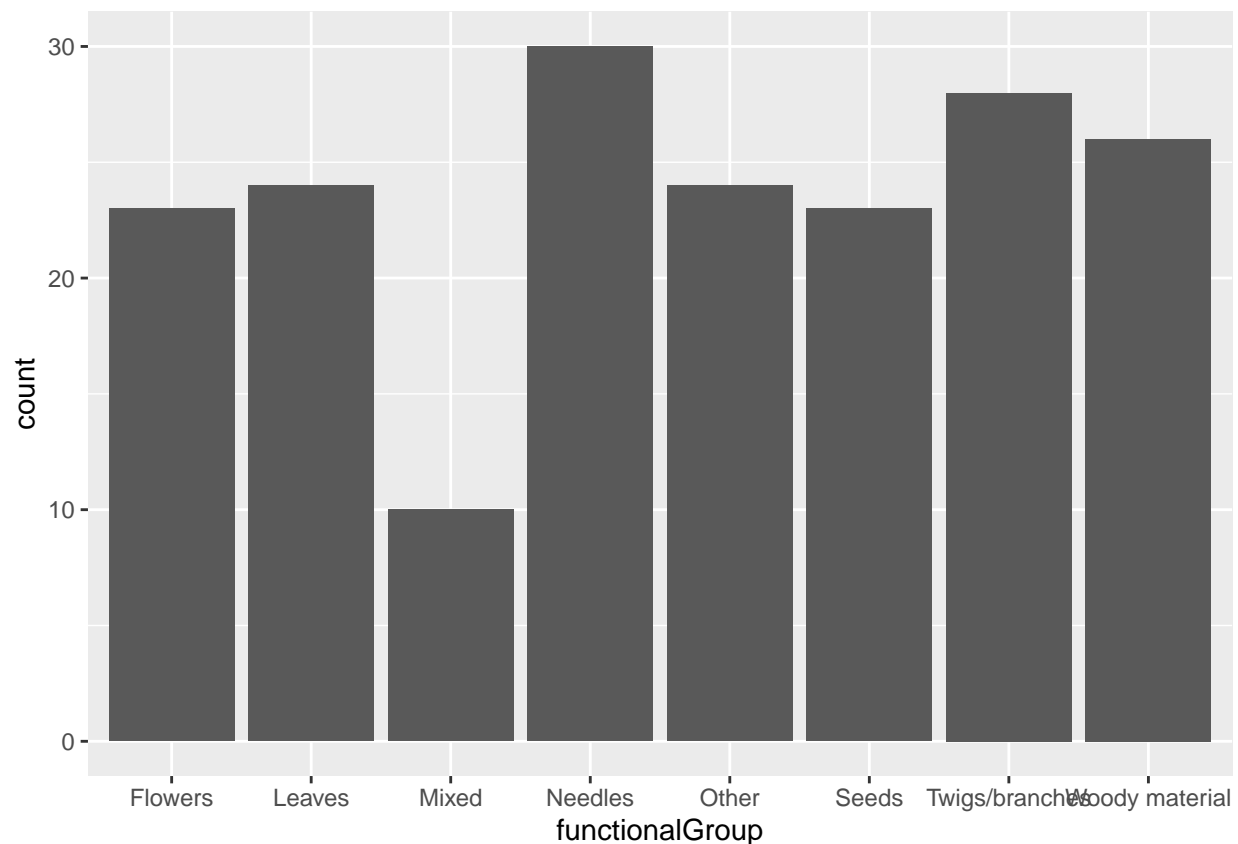
```
summary(df_unique) #summary of plots sampled at Niwot Ridge
```

```
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061
##          1          1          1          1          1          1          1          1
## NIWO_062 NIWO_063 NIWO_064 NIWO_067
##          1          1          1          1
```

Answer: Unique shows the sum of plots at the Niwot Ridge area, however summary shows how many individual counts there are of Niwot Ridge.

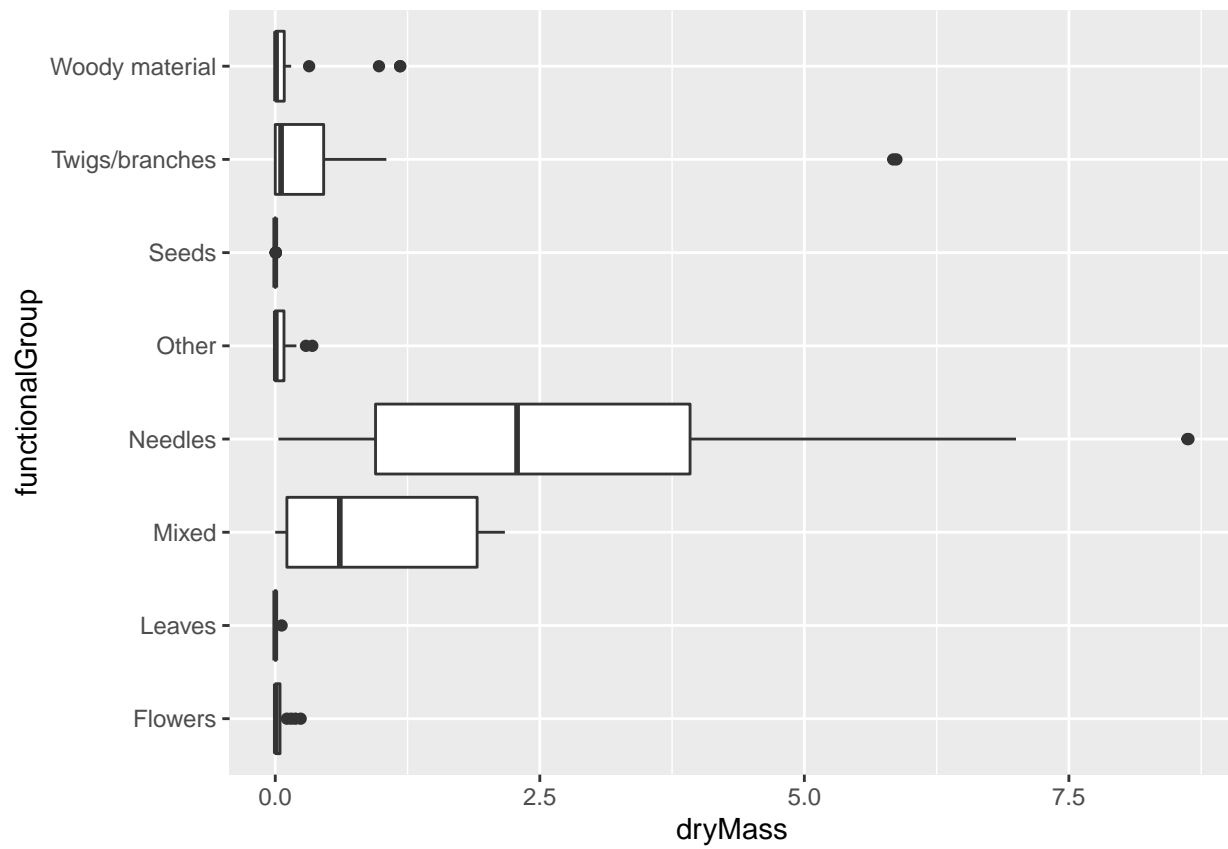
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

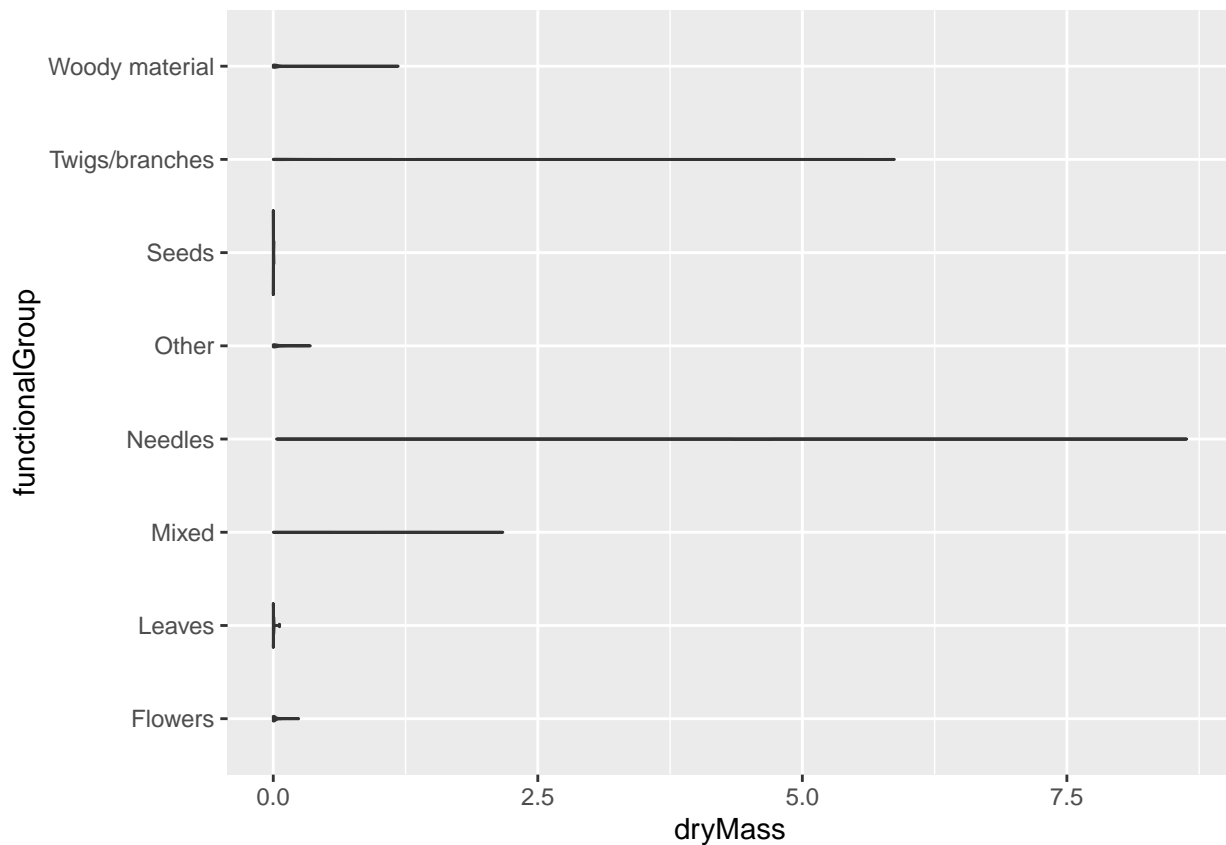


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of `dryMass` by `functionalGroup`.

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

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



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

Answer: The box plot is more effective than the violin plot because it shows the mean as well as the quartile range. The violin plot alternatively just shows the entire range of results which is not as indicative for what is commonly observed.

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

Answer: The litter taht tends to have the highest biomass is needles.