Assignment 3: Data Exploration

Sam Tolbert

Fall 2024

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 Canvas.

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. Load necessary packages (tidyverse, lubridate, here), check your current working directory 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.

```
library(tidyverse) #getting libraries
library(lubridate)
library(here)

Neonics <- read.csv ( #openning and naming files
file = here('./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv'),
stringsAsFactors = TRUE
)

Litter <- read.csv( #opening and naming
```

```
  file= here('./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: They are an insecticide that binds to the nerves of insects, causing them ot overstimulate and die. They are applied as a "drench" meaning 95% of what is applied ends up in the soil and eventually the water supply. It has been shown to have a negative effect on bees, who provide pollination and a vital part of plant life, as well as to vulnerable human population.

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: Litter and woody debris is a vital step in natural wildfire cycles. The monitoring and maintenance of woody debris can help us regulate controlled burns and also is important in understanding when natural burns are most needed and most productive to forest health.

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. 2. 3.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) #4623 observations of 30 variables
```

```
## [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? [Tip: The `sort()` command is useful for listing the values in order of magnitude...]

```
Neonics_summary<- summary(Neonics$Effect)

Neonics_HightoLow <-sort(Neonics_summary, decreasing=TRUE)

print(Neonics_HightoLow) #fancy way
```

```
##       Population        Mortality         Behavior Feeding behavior
##             1803             1493              360              255
##     Reproduction      Development        Avoidance         Genetics
##              197              136              102               82
##        Enzyme(s)           Growth       Morphology    Immunological
##               62               38               22               16
##     Accumulation     Intoxication     Biochemistry          Cell(s)
##               12               12               11                9
##       Physiology        Histology       Hormone(s)
##                7                5                1
```

```
sort(summary(Neonics$Effect), decreasing= TRUE) #less fancy way; is there a
```

```
##       Population        Mortality         Behavior Feeding behavior
##             1803             1493              360              255
##     Reproduction      Development        Avoidance         Genetics
##              197              136              102               82
##        Enzyme(s)           Growth       Morphology    Immunological
##               62               38               22               16
##     Accumulation     Intoxication     Biochemistry          Cell(s)
##               12               12               11                9
##       Physiology        Histology       Hormone(s)
##                7                5                1
```

```
#best practices for how to do this?
```

Answer: Most common effect is population, which is vital when understanding the systemic effects an insecticide could have on a species or collection of species. How much does it kill is an important question.

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: Explore the help on the `summary()` function, in particular the `maxsum` argument...]

```
NeonicsSortedSpecies<- (sort(
  summary(Neonics$Species.Common.Name, maxsum=7) #maxsum=7 instead of 6
  , decreasing= TRUE #because otherwise "other" as the most common crowds out
  ) #number 6
)

print (NeonicsSortedSpecies)
```

```
##               (Other)             Honey Bee        Parasitic Wasp
##                  3083                   667                   285
## Buff Tailed Bumblebee   Carniolan Honey Bee            Bumble Bee
##                   183                   152                   140
##      Italian Honeybee
##                   113
```

Answer: Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee are most commmon species because they are all pollinators. If these are killed by the insectcide it will negatively effect the whole ecoystem, meaning they are vital to study

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? [Tip: Viewing the dataframe may be helpful...]

```
class(Neonics$Conc.1..Author.)
```

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

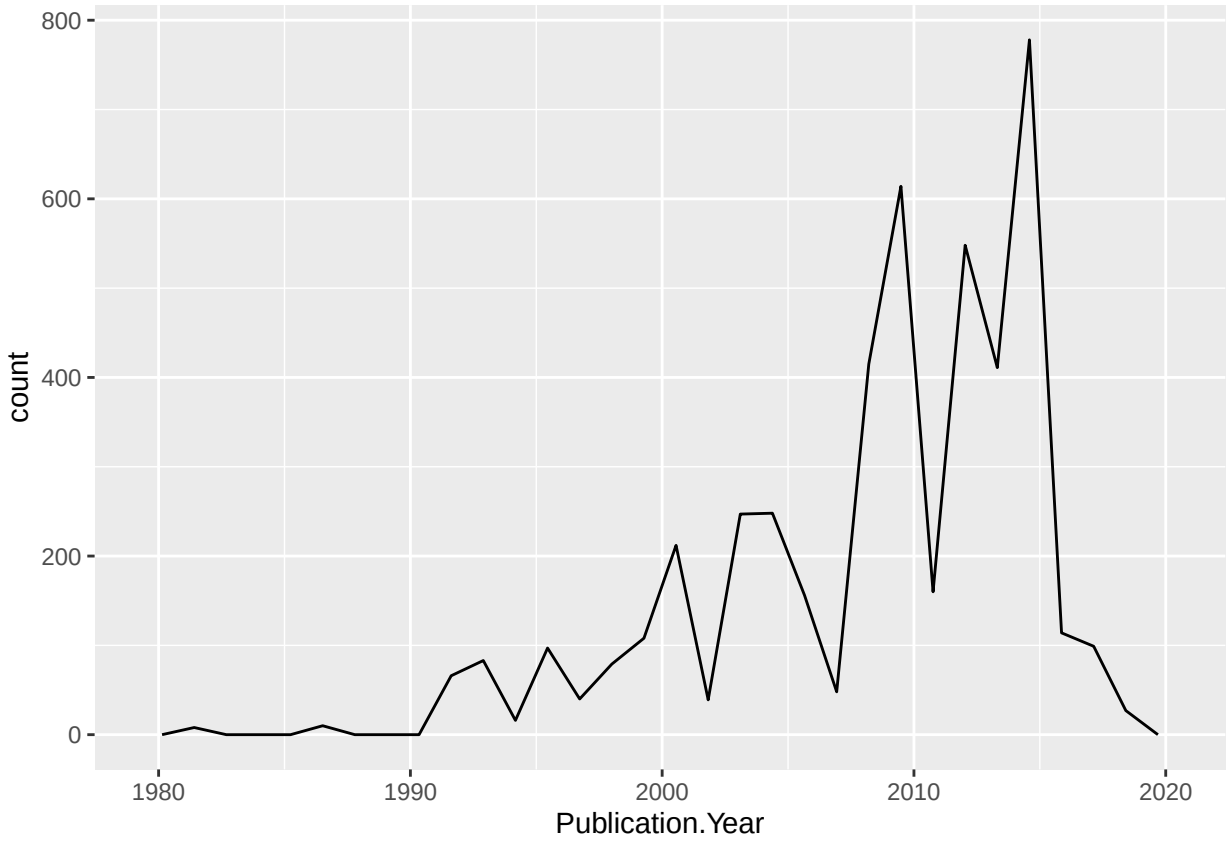
Answer: It is a factor. It is not numeric because mix of quantitive and qualitivate

Explore your data graphically (Neonics)

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

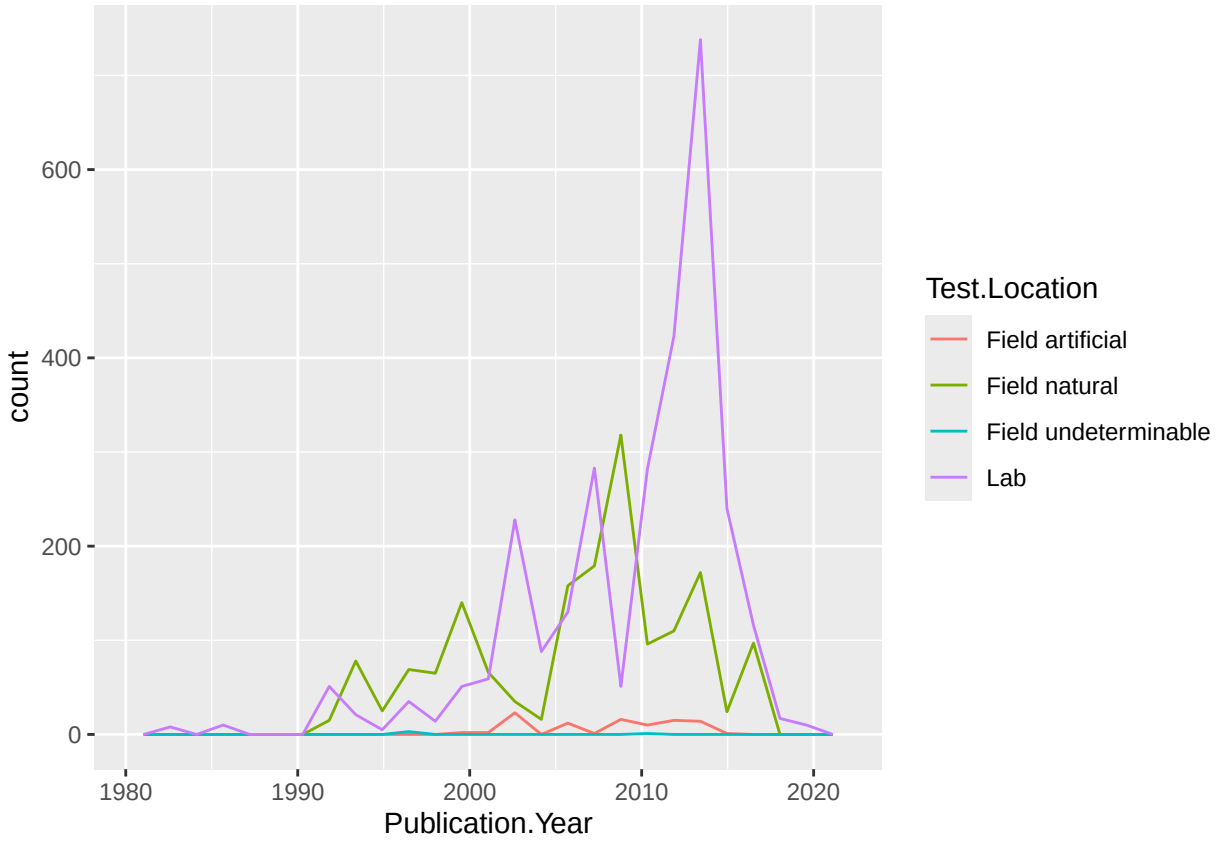
```
Neonics_Frequency<-ggplot(Neonics) +
  geom_freqpoly (aes(x=Publication.Year)
)

print(Neonics_Frequency)
```

```
## `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.

```
Neonics_Frequency<- ggplot(Neonics) +
  geom_freqpoly (aes(x=Publication.Year, color= Test.Location), bins=25
)

print(Neonics_Frequency)
```



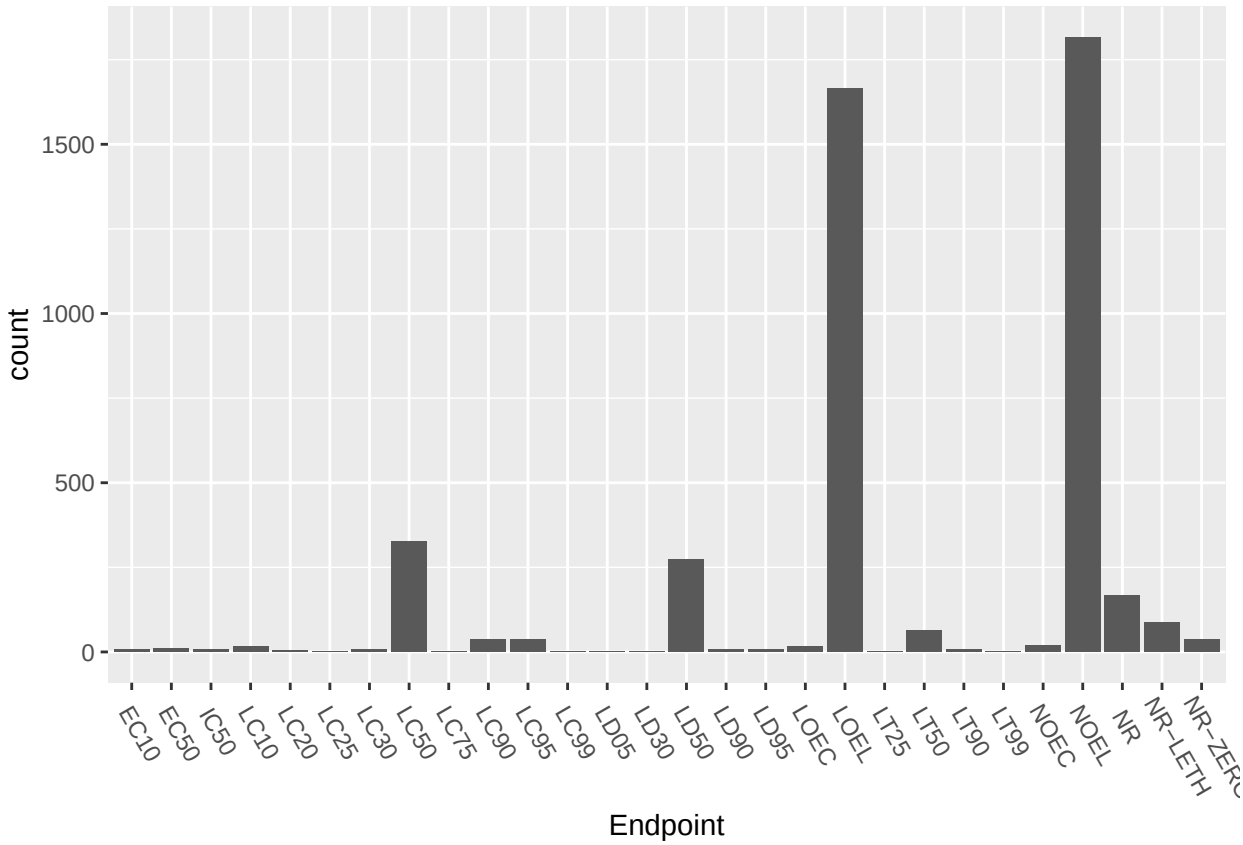
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer:The most common test location recently is Lab, however there are peroids (~1992-2002, 2008-2010) were "field natural" was 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.

[**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...]

```
Neonics_Bar <- ggplot(Neonics, aes(x=Endpoint)) + geom_bar() + theme(axis.text.x = element_text(angle =

print(Neonics_Bar)
```



Answer:The two most common endpoints are NOEL, defined as 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) and LOEL, defined as Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC).

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.

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

library(lubridate)

Litter$collectDate<- ymd(Litter$collectDate)

unique(Litter$collectDate) #only two dates, 2018-08-02 and 2018-08-30
```

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

13. Using the `unique` function, determine how many different 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 NIWO_047 NIWO_051
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

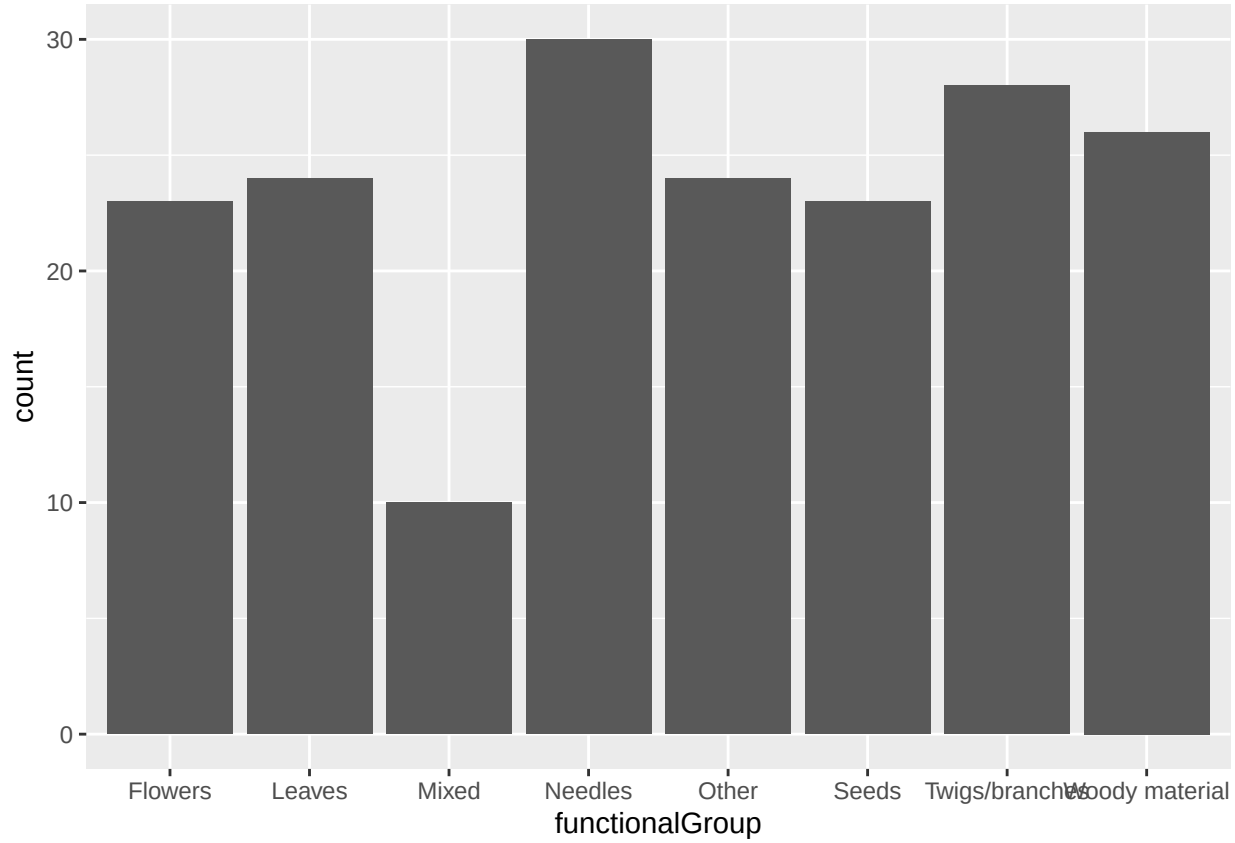
```
summary(Litter$plotID)
```

```
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061
##       20       19       18       15       14        8       16       17
## NIWO_062 NIWO_063 NIWO_064 NIWO_067
##       14       14       16       17
```

Answer: With unique() it will tell me the number of unique classes that I ask for, but summary of that class will list each one out (if there's not too many I can count) but also it will list the number of observations within each class

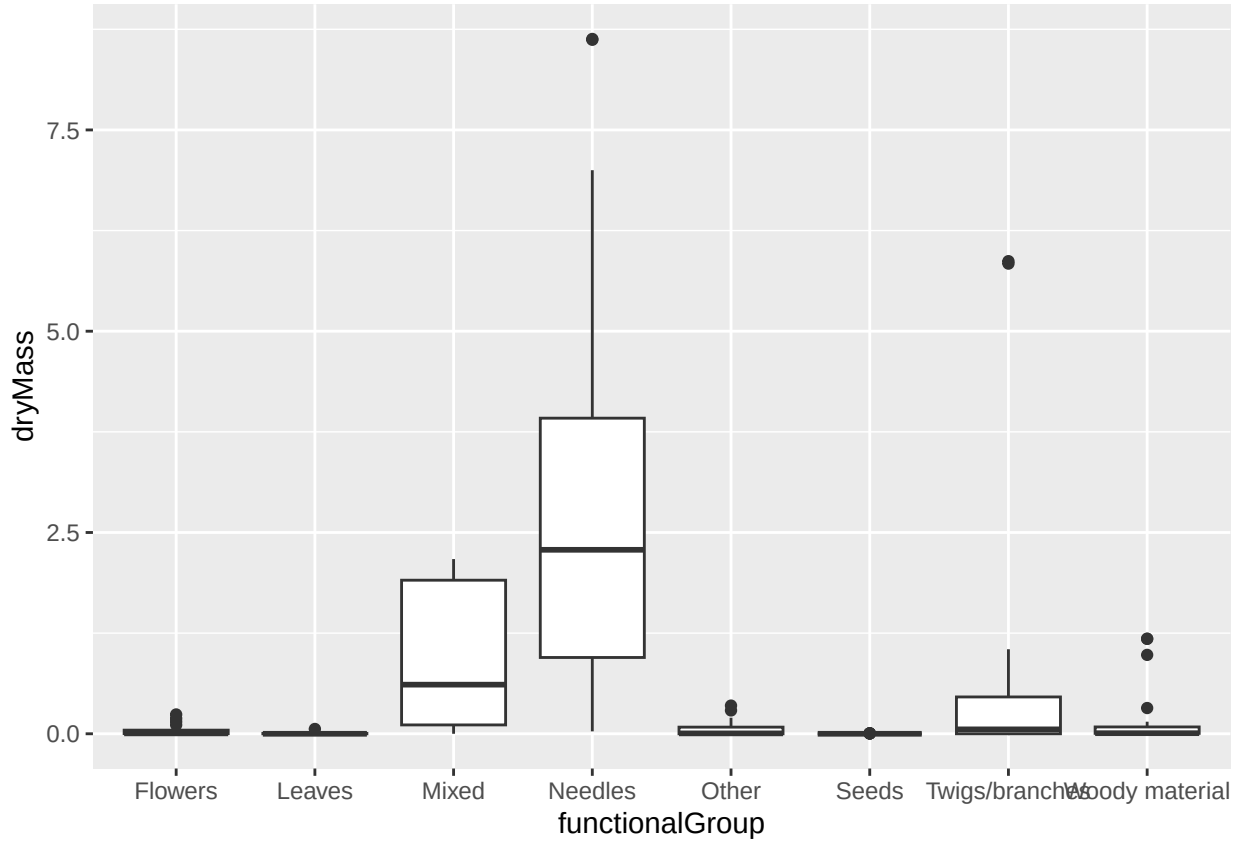
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
Litter_functionalGroup <- ggplot(
  data=Litter, aes(x=functionalGroup)) +
  geom_bar()

print(Litter_functionalGroup)
```

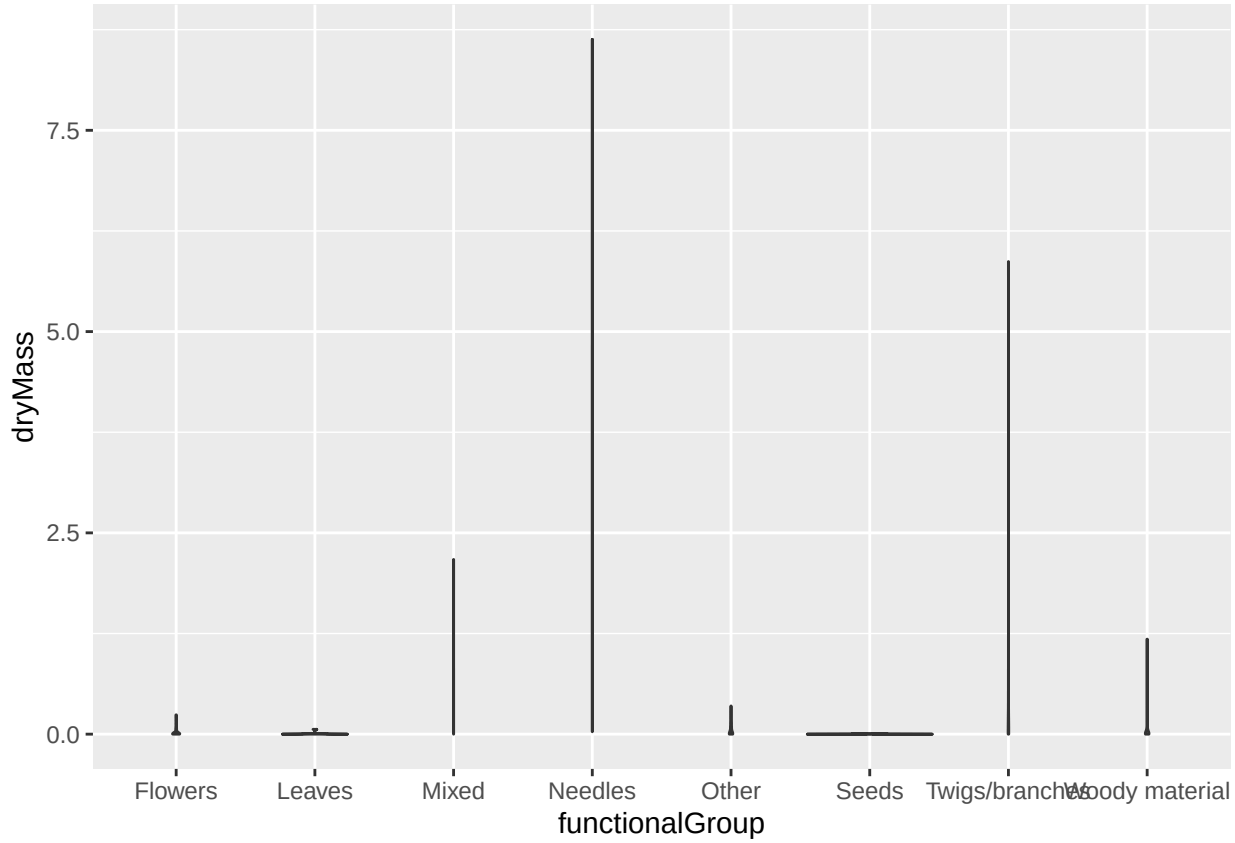


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
Litter_dryMassBox <- ggplot(
  data=Litter, aes(x=functionalGroup, y= dryMass)) +
  geom_boxplot()

print(Litter_dryMassBox)
```




```
Litter_dryMassFiddle <- ggplot(
  data=Litter, aes(x=functionalGroup, y= dryMass)) +
  geom_violin()

print(Litter_dryMassFiddle)
```



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

> Answer: The box plot is more effective because it more effectivley conveys the range, in this case using y axis to try to convey the size of each range instead of the x. The violin plot uses the x axis instead of to convey the size of each indiviudal part of the range, and in this case the granularity doesn't create a ton of clarity.

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

> Answer: Needles by a long shot.