Assignment 3: Data Exploration

Brock Keller

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.

```r
#install.packages("tidyverse") #commenting these out after I installed them
#install.packages("lubridate")
#install.packages("here")

library(tidyverse) #now loading them in
library(lubridate)
library(here)

getwd() #checking working directory
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```
Neonics <- read.csv(
 file = here("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"),
  stringsAsFactors = TRUE)

#Neonics #calling on it and commented out

Litter <- read.csv(
  file = here("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv"),
  stringsAsFactors = TRUE)

#Litter #calling on it and commented out
```

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: Given that neonicotinoids are a class of pesticide, we are interested in their ecotoxicological effect on the target species as well as on \ other species of insect. A wide range of insects could be affected because chemicals spread through the environment through the air, agricultural washoff, or other mechanisms that can result \ in persistence in the environment or biomagnification. For example, a toxin can pass from a bug that didn't die from it into the bird that eats that bug.

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: Studying the biomass of litter and woody debris in a forest would be important assessing load for wildfires. This type of study probably informs control burns and helps identify areas of concern and response strategies. It could also be interesting data for ecological studies in terms of what forest floor species there could be, overall ecosystem health, and overall productivity.

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. In densely vegetated plots, trap placement is random while in less densely vegetated plots, traps are placed strategically around plants/trees of interest. 2. Some traps are on the ground while others are elevated above it. 3. Frequency of sampling depends on the type of trees in the forest.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#str(Neonics) #gives the structure of the dataset. Tells me it is 4632 observations of 30
#variables (30 columns by 4632 rows) - commented out
```

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...]

```
effects <- (Neonics$Effect) #isolating the column I want
summary(effects) #summarizing that 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
```

```
class(effects) #couldn't get it to sort by magnitude so I saw here it is stored as a factor
```

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

```
levels(effects) #seeing what the levels of the factor are
```

```
##  [1] "Accumulation"     "Avoidance"        "Behavior"         "Biochemistry"
##  [5] "Cell(s)"          "Development"      "Enzyme(s)"        "Feeding behavior"
##  [9] "Genetics"         "Growth"           "Histology"        "Hormone(s)"
## [13] "Immunological"    "Intoxication"     "Morphology"       "Mortality"
## [17] "Physiology"       "Population"       "Reproduction"
```

```
summary(effects, maxsum = 6)
```

```
##       Population        Mortality         Behavior Feeding behavior
##             1803             1493              360              255
##     Reproduction          (Other)
##              197              515
```

Answer: The 6 most commonly studied effects are: population, mortality, behavior, feeding behavior, repruction, and other. It makes sense that these \ effects have the most interest because they relate most directly to the size, health, and basic functions of a species. In the context of a pesticide, we \ want to know whether an insect will come into contact (especialy via eating) with the chemical.

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…]

```
species <-(Neonics$Species.Common.Name)
#summary(species) - commenting out
summary(species, maxsum = 6)
```

```
##            Honey Bee        Parasitic Wasp Buff Tailed Bumblebee 
##                  667                   285                   183 
##   Carniolan Honey Bee            Bumble Bee               (Other) 
##                  152                   140                  3196 
```

Answer: The six most commonly studied species are, in order: honey bee, parasitic wasp, buff tailed bumblebee, carniolan honey bee, bumblebee, and other. All but the wasp \ and "other" are bee species and wasps are still relevant because they are predatory to bees. It makes sense that there is more interest in bees \ because they are polinators and essential to ecosystem health- a lot of people care a lot about them.

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.) #says it is a factor
```

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

```
#view(Neonics$Conc.1..Author.) #commenting out for knit
```

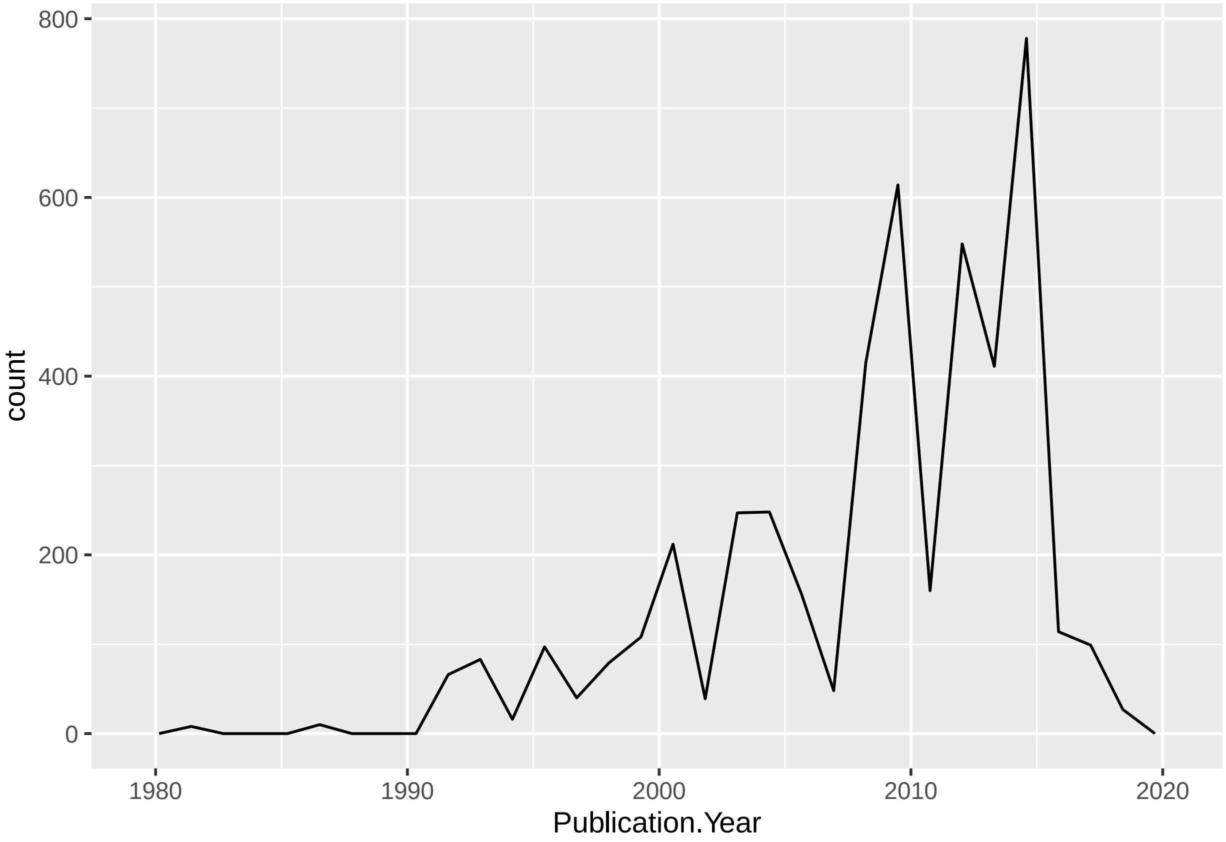
Answer: Conc.1..Author is stored as a factor. It is not numeric because not every attribute has associated data or \ the data is in a different format. "NA" is not a number and the first rule \ of ecotoxicology is that there is no such thing as 0, only <LOD!

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))
```

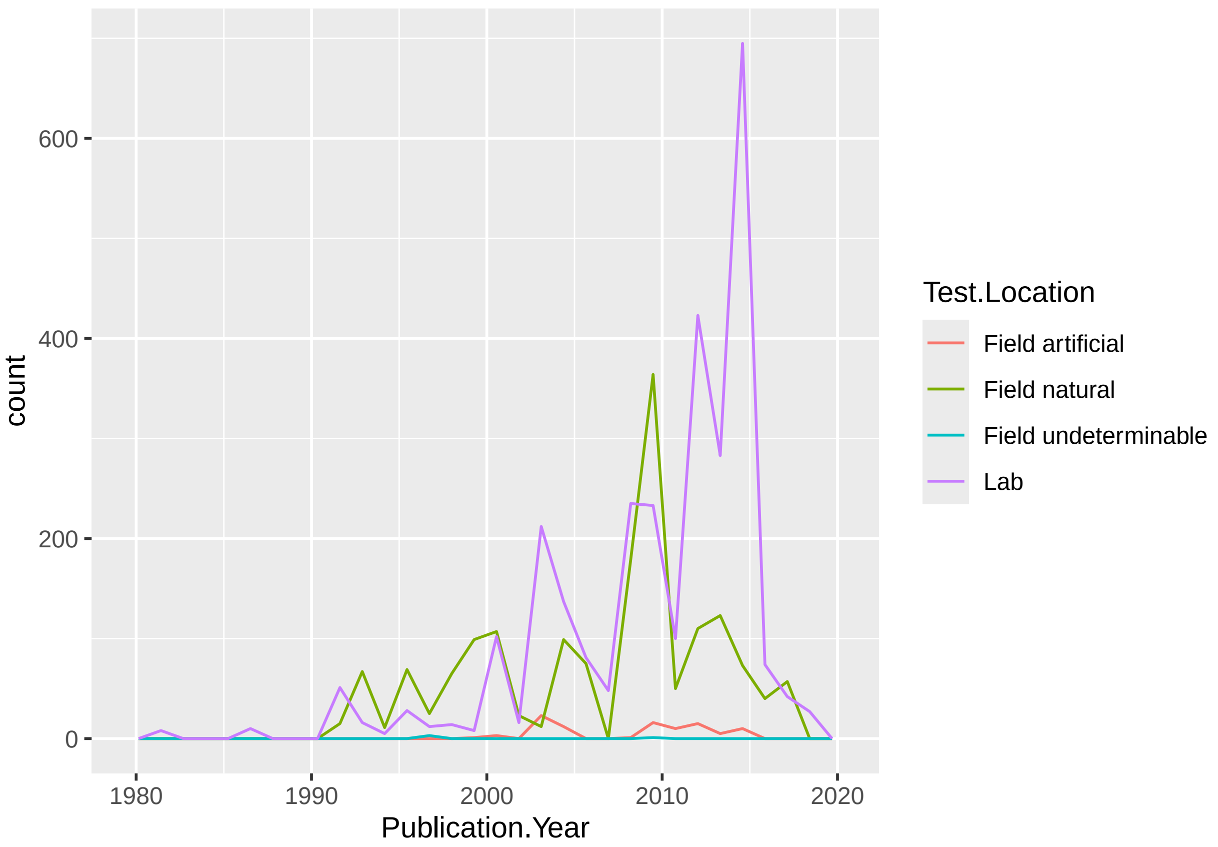
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

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

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



```r
#location, shows up with key in new plot

summary(Neonics$Test.Location) #checking for accuracy
```

```
##     Field artificial        Field natural Field undeterminable 
##                   96                 1663                    4 
##                  Lab 
##                 2860 
```

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

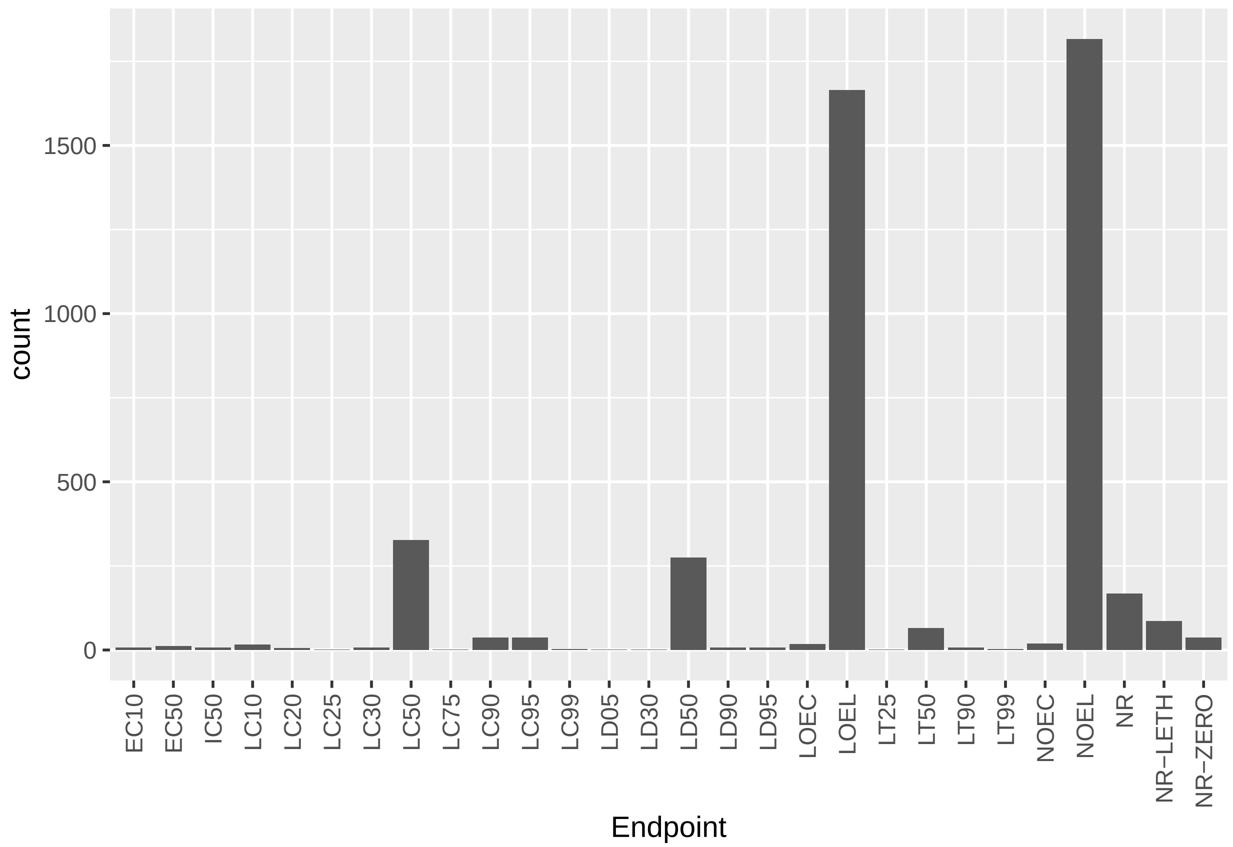
> Answer: Lab and Field natural are by far the most common. Starting around 2000 the number of lab and field natural test locations increase dramatically in frequency. Field artificial shows up a few times from about 2002 to 2016, while Field undeterminable is very rarely seen.

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...]

```r
#summary(Neonics$Endpoint, maxsum = 2)

ggplot(Neonics) +
  geom_bar(aes(x = Endpoint)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: The two most common are NOEL and LOEL, corresponding to the No Observable Effect Level and \ Lowest Observable Effect Level respectively. Both refer to the doses at which the described response is elicited.

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)
```

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

```
#says this is a factor, not a date.

Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d") #telling R directly
#that collectDate is a date, not a factor and setting the format
#Litter$collectDate #calling on it to check change - commented out

class(Litter$collectDate) #verifying change in data class
```

```
## [1] "Date"
```

```
august_2018_dates <- Litter$collectDate[format(Litter$collectDate, "%Y-%m") == "2018-08"]
#setting object as only dates that have 2018 for year and 08 for month
#august_2018_dates - commented out

unique(august_2018_dates) #using unique function on my narrowed august dates object
```

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

```
#so this shows that the only two dates for collection in August 2018 were the 2nd and the 30th
```

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`?

```
summary(Litter$namedLocation) #these two things are telling me the same thing in just a
```

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

```
#slightly different way.
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
```

```
plots <- as.character(Litter$plotID) #converting to a character instead of as a factor
class(plots)
```

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

```
different_plots <- unique(plots) #defining the unique plots
different_plots
```

```
##  [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(different_plots) #having R summarize how many unique plots there are from the character data
```

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

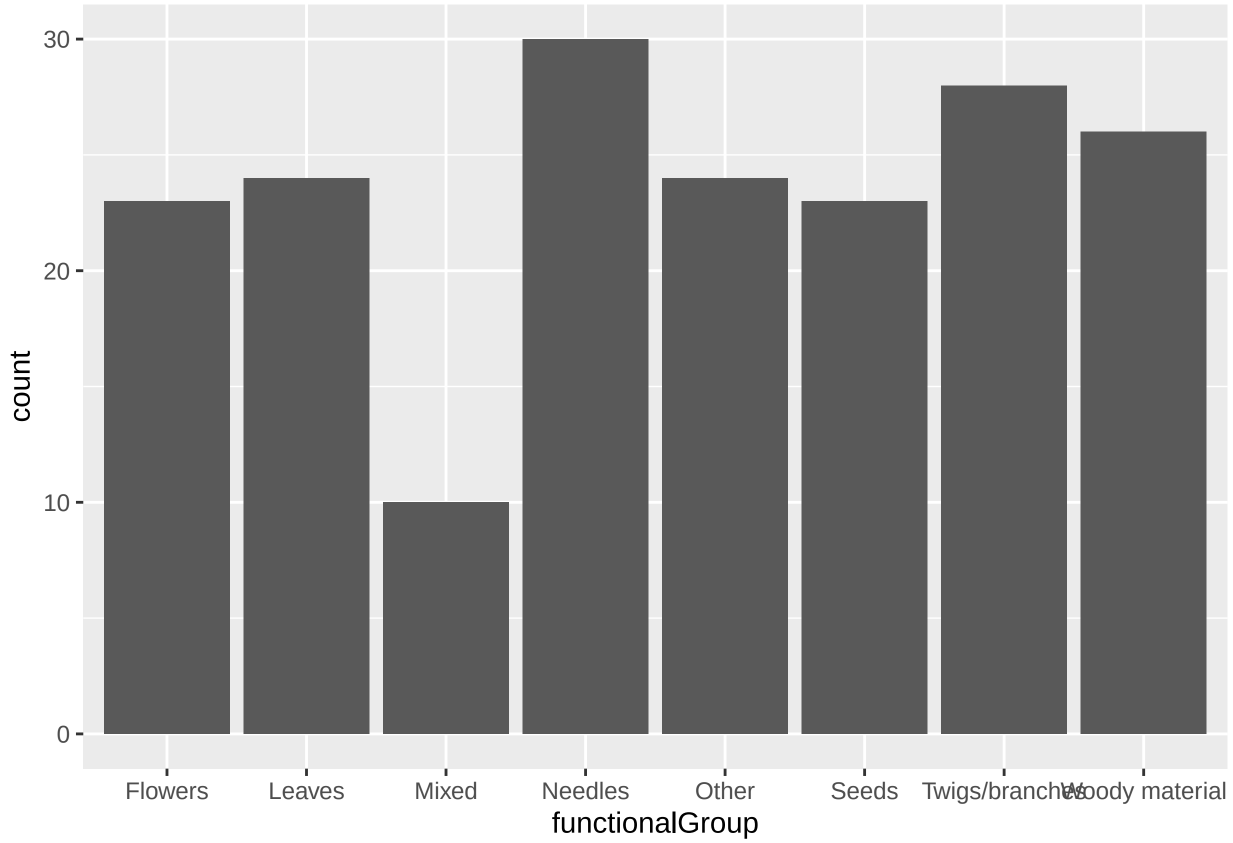
```
summary(as.factor(Litter$plotID))
```

```
## 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: The unique() function shows that there are 12 individual plots sampled at Niwot Ridge. \ Unique() provides only the plots themselves while summary() shows me the associated data for each plot.

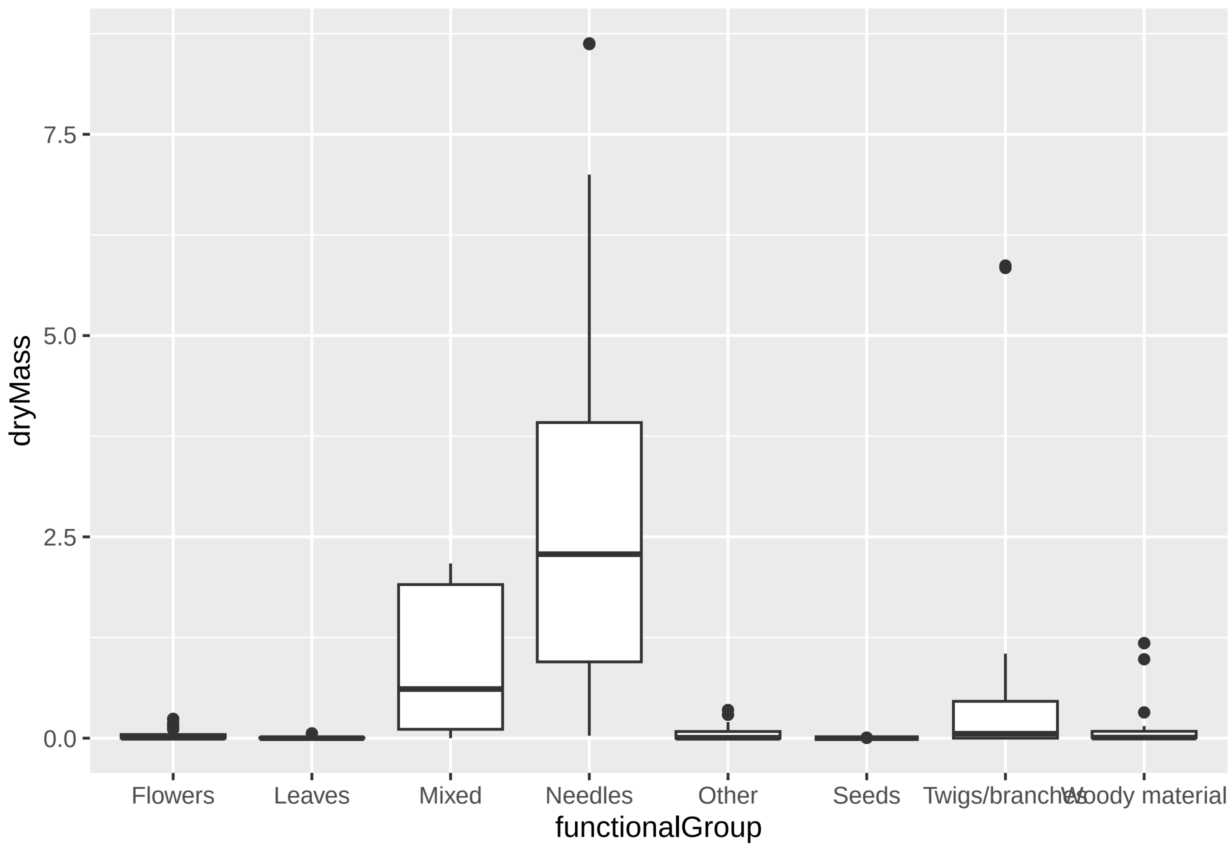
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

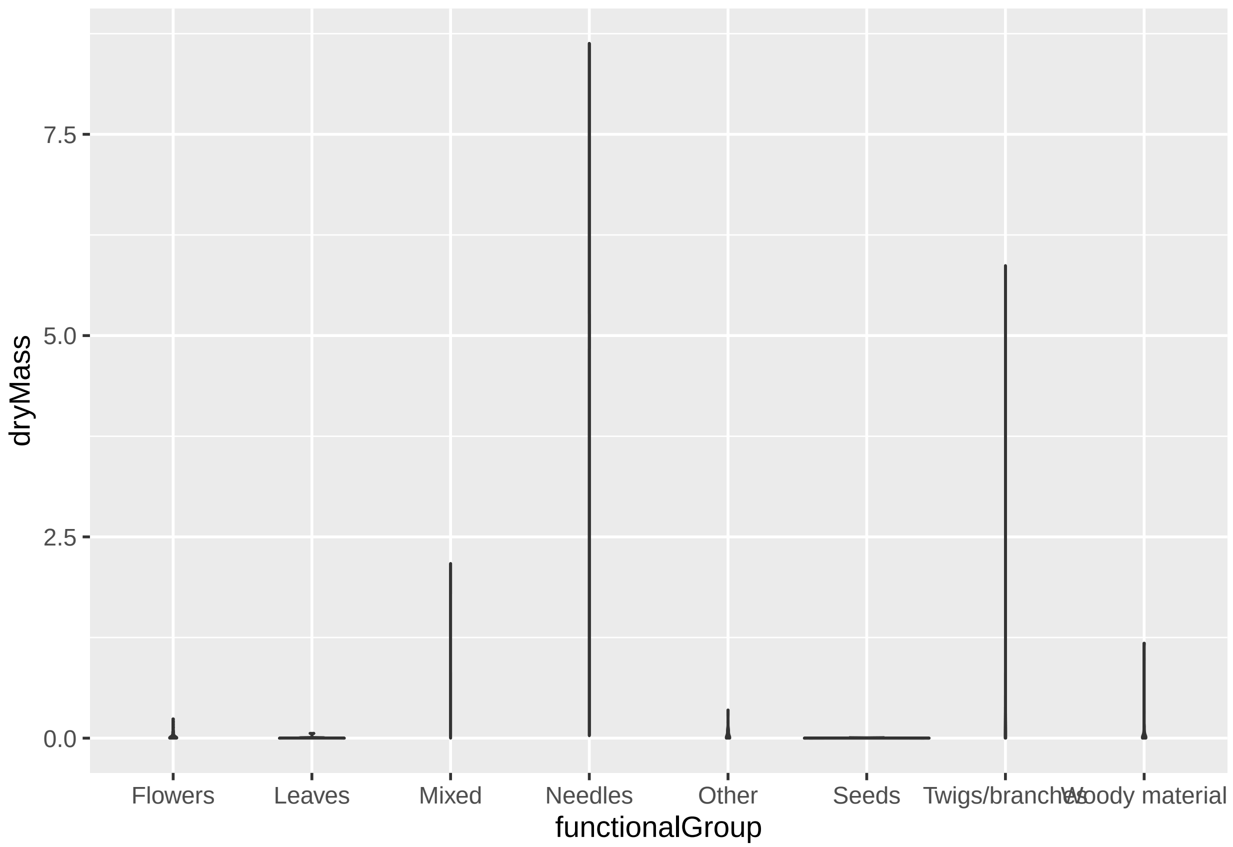


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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



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



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 most of the data have very low values despite there being several outliers and wide range in values. The violin plots are not 'zoomed in' enough to effectively see the quantile distributions.

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

> Answer: Needles tend to have the highest biomass at the Niwot Ridge sites.