Assignment 3: Data Exploration

Jingze Dai

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.

```
# loading packages
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(here)
```

```
## here() starts at /home/guest/ENV872/EDE_Fall2024
```

```
# checking work directory
getwd()
```

```
## [1] "/home/guest/ENV872/EDE_Fall2024"
```

```
# we can see that the current work directory is /home/guest/ENV872/EDE_Fall2024

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

Litter <- read.csv(
  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: Studying the ecotoxicology of neonicotinoids on insects is important because of the potential negative impacts that neonicotinoids may bring to non-target insect species, especially those beneficial ones such as bees. According to Environment America, neonicotinoids targets the nervous systems of bees, paralyzing adult bees and harming baby bees' brains. Therefore, neonicotinoids can harm insects that are critical for pollination, potentially damaging the entire ecosystem. (source: "3 ways neonic pesticides are harming bees", SAVE THE BEES, Published April 19, 2024, accessed September 24, 2024. https://environmentamerica.org/articles/3-ways-neonic-pesticides-are-harming-bees/#:~:text=Neonics%20attack%20the%20bee's%20central,and%20the%20colony%20could%20collapse)

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: According to U.S. Department of Agriculture, studying woody debris and other forest litter is important because of their crucial role in carbon budgets and nutrient cycling. Studying these give us insights on the soil fertility of the forest thus these litter act as an indicator of the forest health. (source: Donna B. Scheungrab, Carl C. Trettin, Russ Lea, Martin F. Jurgensen, "Woody Debris", Forest Service, U.S. Department of Agriculture, published in 2000, accessed September 24, 2024. https://research.fs.usda.gov/treesearch/20001#:~:text=Woody%20debris%20is%20an%20important,influencing%20water%20flows%20and%20sediment)

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. Litter sampling sites that targets forested tower airsheds have 20 40m x 40m plots. Those targeting low-saturated vegetation have 4 40m x 40m tower plots plus 26 20m x 20m plots. In total, there are 1 to 4 trap pairs per plot. 2. Traps are placed randomly within the grid cell for sites with >50% aerial cover of woody vegetation > 2m. Traps are placed in a targeted way for areas with <50% cover of woody vegetation, and are strategically placed beneath heterogeneously distributed and patchy vegetation. 3. Ground trap sampling frequency is once per year. Elevated traps are sampled once per week in deciduous forest sites during senescence and once every one to two months in evergreen sites.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
# use the dim function
dim(Neonics)
```

```
## [1] 4623   30
```

```r
# there are 4623 entries and 30 variables
```

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…]

```r
# using sort command to organize the rank of the summary of "Effect" column
sort(summary(Neonics$Effect))
```

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

Answer: The most common effects being studied are mortality (1493 entries) and population (1803 entries). These two are of interest because the effect on mortality has a direct relationship with whether the tested insect is resistent to the toxicity of neonicotinoids, and by assessing the effect on population, the impact of neonicotinoids on population of different types of insect can be known, which indicates the ecosystem's response to the pesticide.

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…]

```r
# maxsum is an integer indicating how many levels should be shown for factors
# we are interested in the top six most studied species
# therefore we can set maxsum = 7 (6 species plus 'others') in this case
summary(Neonics$Species.Common.Name, maxsum = 7)
```

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

Answer: We can see that the top six most commonly studied species are honey bee (667 times), parasitic wasp (285), buff tailed bumblebee (183), Carnioan honey bee (152), bumble bee (140), and Italian honeybee (113). We can see that they all belong to bees or wasps. The bees are pollinators and if we regard parasitic wasps as a biogical control insect, all of the six speicies are beneficial insects. Moreover, they are all sensitive to environmental stressors, especially pesticided. Studying their response to neonicotinoids can provide early warnings about the health level of the ecosystem.

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…]

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

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

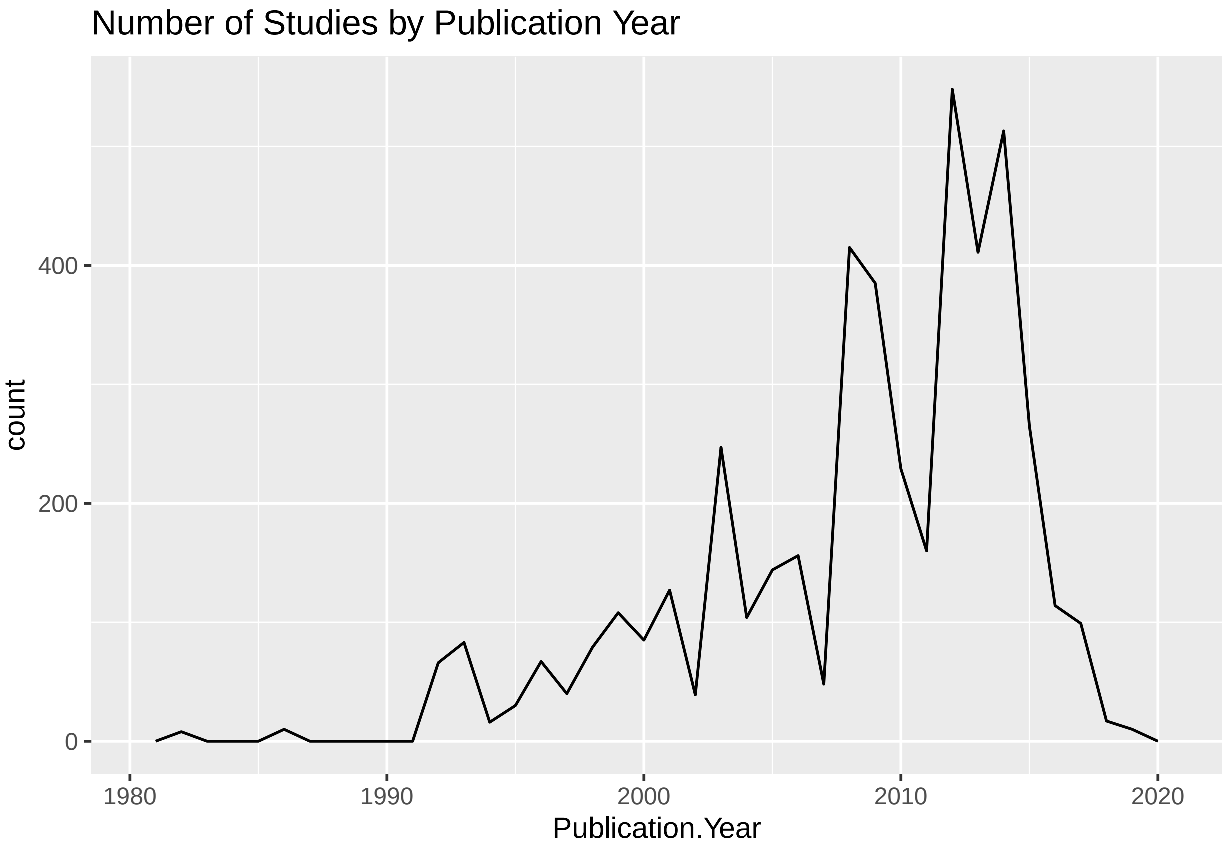
Answer: By taking a look at the dataset, I saw that concentrations are in different units, such as AI lb/acre, %, fl oz/gal and many others. It would be then meaningless, or even misleading, to record the concentrations in numeric when the units are totally different. Moreover, some concentrations have '/' behind them and some have the value of NR. These conditions also make it inappropreiate to label them as numerals.

Explore your data graphically (Neonics)

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

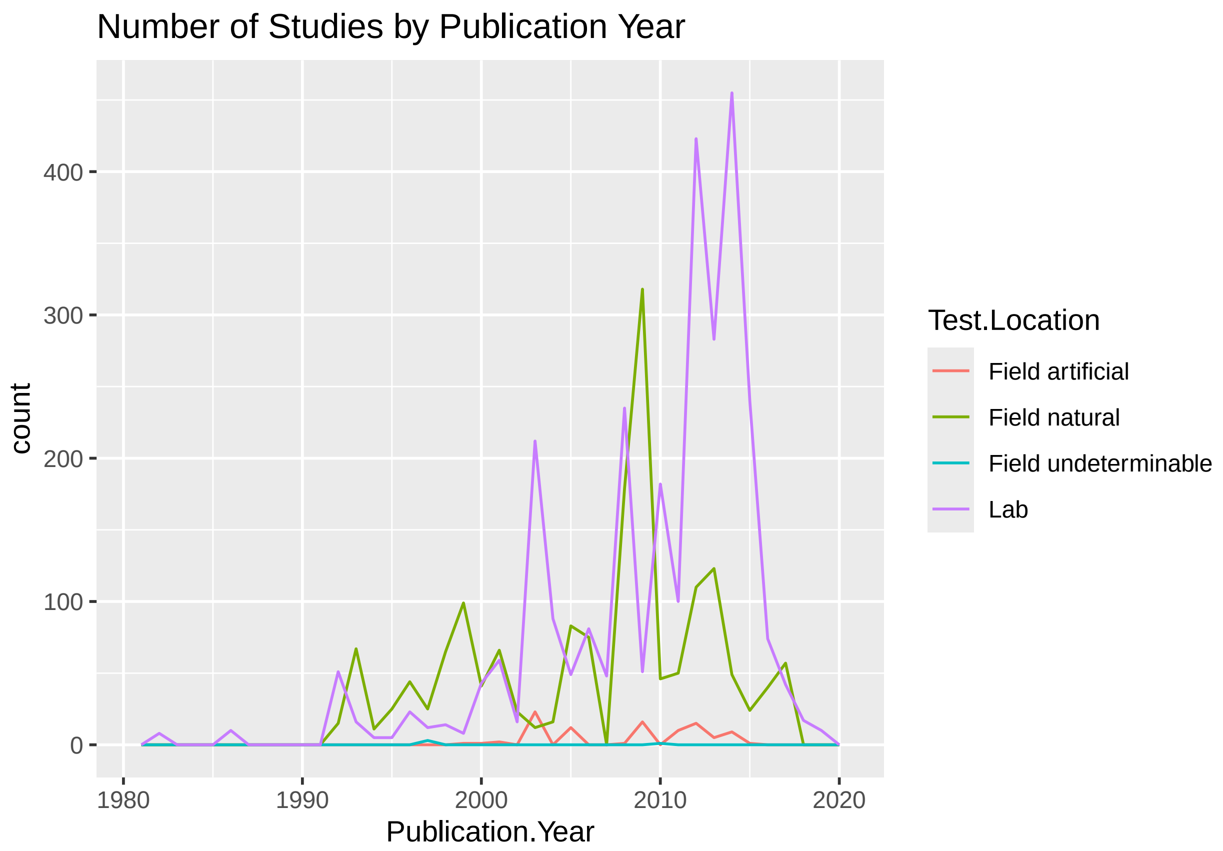
```r
library(ggplot2)

# creating the frequency plot with binwidth = 1
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), binwidth = 1) +
  labs(title = "Number of Studies by Publication Year")
```



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

```r
# color coding by different test locations
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), binwidth = 1) +
  labs(title = "Number of Studies by Publication Year")
```



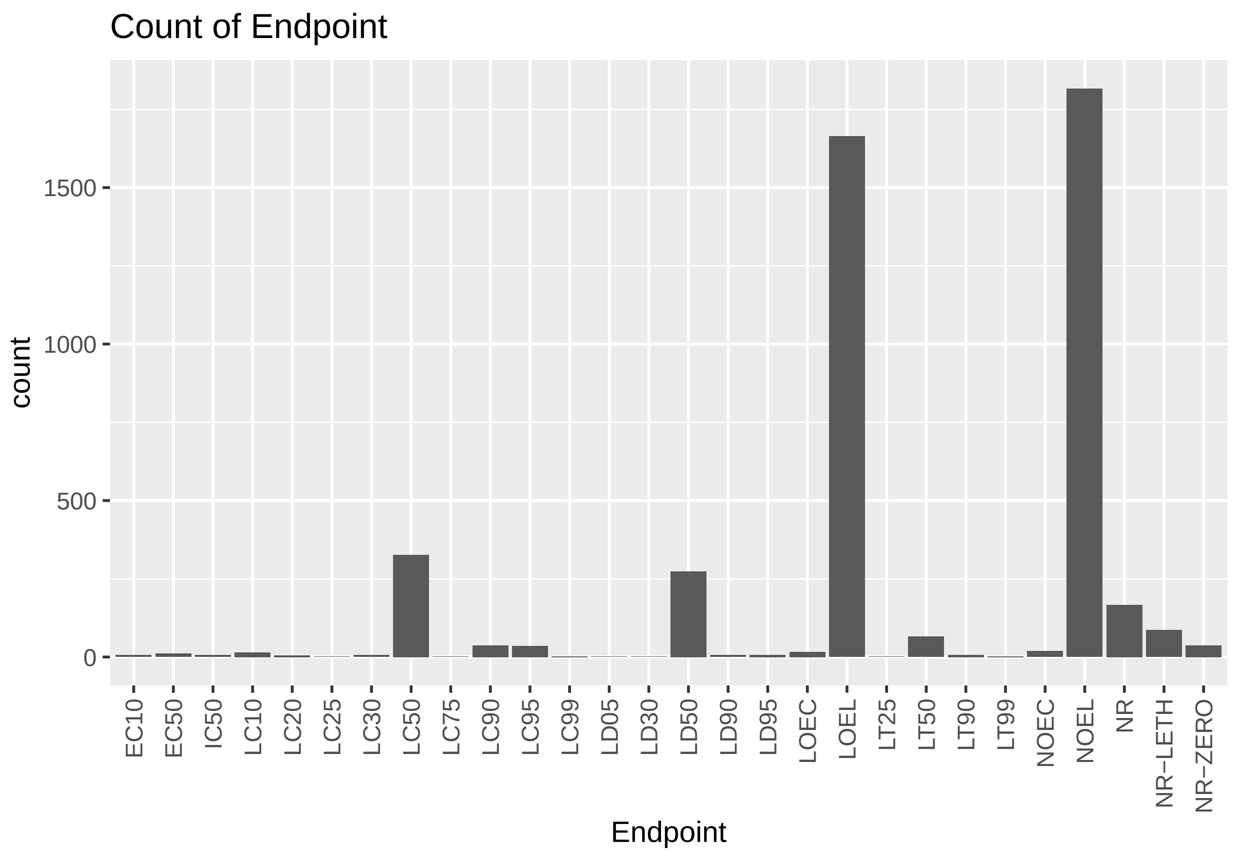
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are field (natural) and labs, and tests conducted in labs are the most for almost all years. Numbers of tests conducted in artificial field or undeterminable field remain low for all years, and they do not differ over time by a lot. However, from time to time, the tests conducted in the two most common locations vary in number and have huge fluctuations over time.

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…]

```
# plotting the bar graph
# labels are rotated 90 degrees for better readability
ggplot(Neonics) +
  geom_bar(aes(x = Endpoint), stat = "count") +
  labs(title = "Count of Endpoint") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: Two most common endpoints are LOEL and NOEL. According to ECOTOX_CodeAppendix, LOEL belongs to the terrestrial database and stands for Lowest Observable Effect Level. NOEL also belongs to the terrestrial database and stands for No Observable Effect Level. LOEL is the lowest dose or concentration producing effects that were significantly different from responses of controls according to statistical tests, and NOEL is highest dose or concentration producing effects not significantly different from responses of controls according to statistical tests. These two endpoints combined can give a safe exposure level for insects.

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"
```

```
# it is a factor, so we need to change it to dates

# the years are all 2018, thus we can simply use as.Date function
# following the original format of YMD
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
```

```
# checking class again
class(Litter$collectDate)
```

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

```
# checking unique dates of sampling
unique(Litter$collectDate[lubridate::month(Litter$collectDate) == 8
                          & lubridate::year(Litter$collectDate) == 2018])
```

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

```
# sampling dates are August 2nd and August 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`?

```
# plotID represents unique plots sampled at Niwot Ridge
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
```

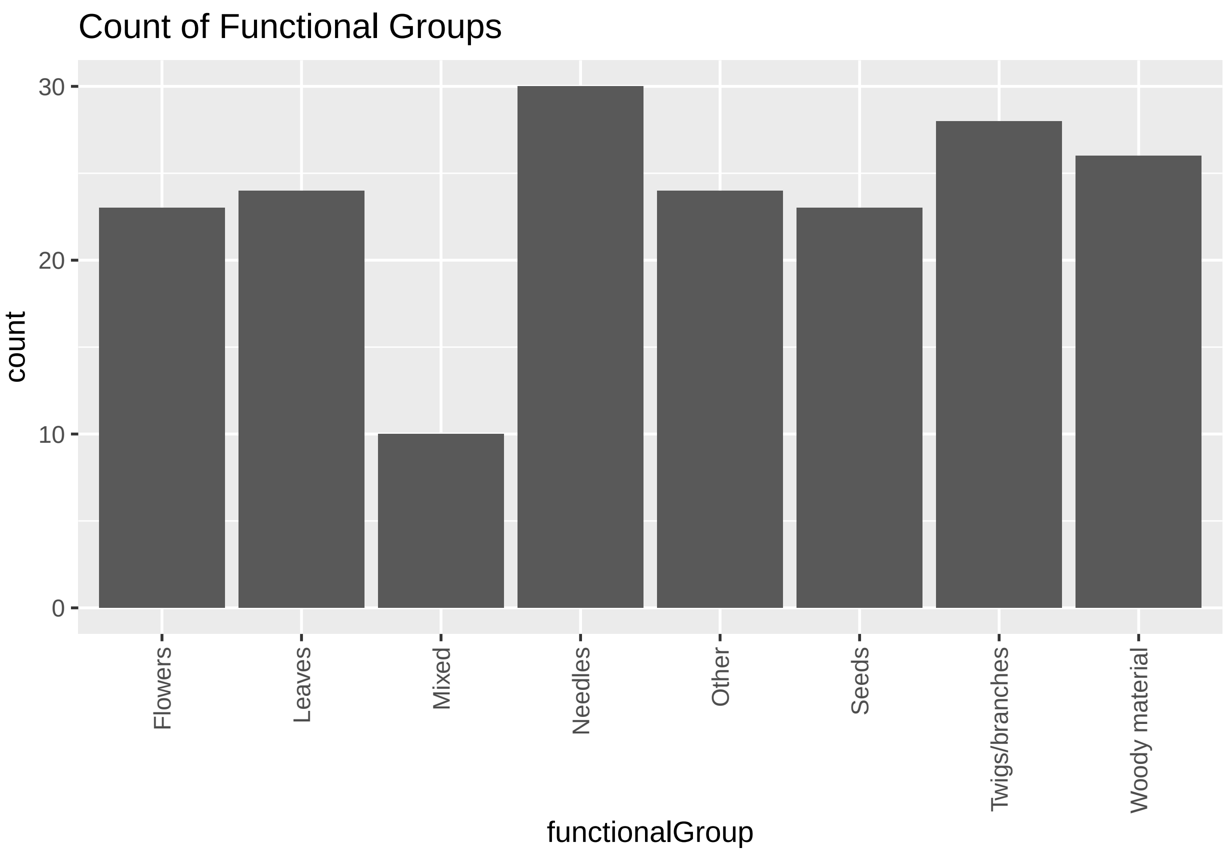
```
# comparing between unique function and summary function
summary(Litter$plotID)
```

```
## 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: The unique function only returns the distinct ID of the plots without counting how many samples there are in each plot, but the summary function gives the count value. Unique function is better in dealing with factors, but summary function can be used for other data types as well.

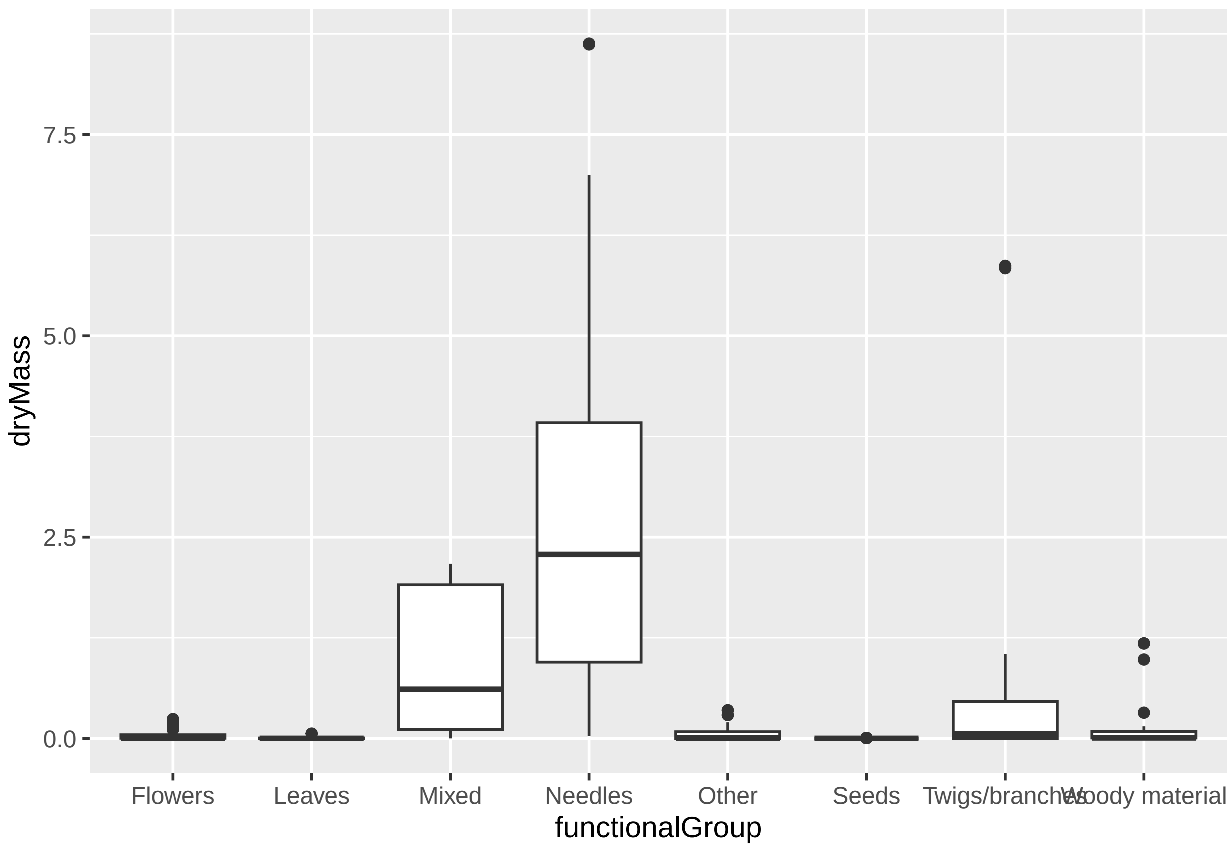
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar(stat = "count") +
  labs(title = "Count of Functional Groups") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

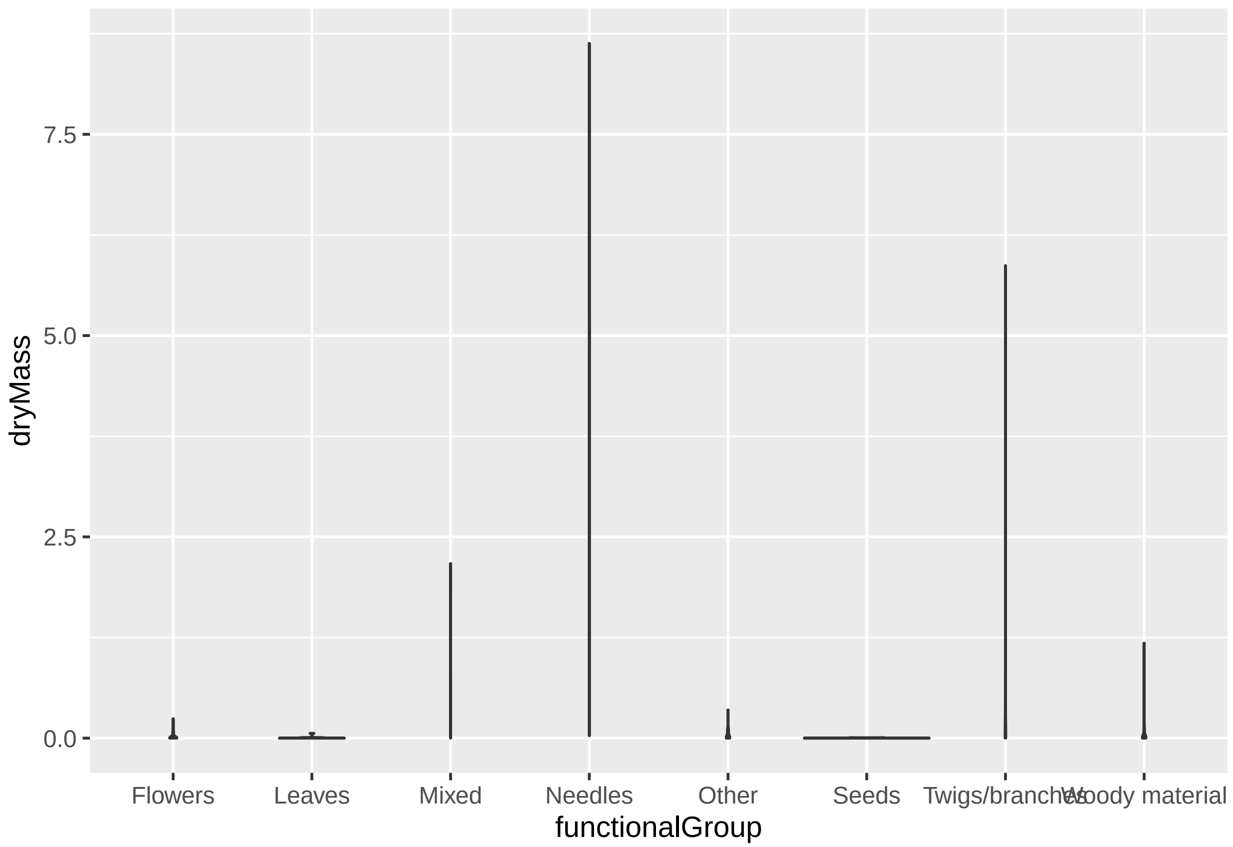


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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



```
# violin plot
ggplot(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 violin plot in this case shows only straight lines, this might be due to the distribution of the dry mass value within each functional group being largely varied. Violin plots accounts for data distribution, but in this dataset the distribution is not useful because of the large variation, thus the boxplot, which only shows the median, quartiles and range, is more effective.

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

> Answer: Needles and mixed, due to their larger mean mass and a higher maximum compared to other types.