Assignment 3: Data Exploration

Yeeun Kim

Spring 2025

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.

```
# Install the necessary packages
library(tidyverse)
library(lubridate)
library(here)

# My current working directory
getwd() # "/home/guest/EDA_Spring2025"
```

```
## [1] "/home/guest/EDA_Spring2025"
```

```r
# Upload the 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 they can negatively affect pollinators by disrupting their foraging, navigation, and reproduction. The pesticides also remain in the environment, which harm the ecosystems. Thus, studying about these topic will help to understand the need for the research on sustainable pesticides.

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 litter and woody debris that falls to the ground in the forests help to understand their importance in ecosystem. They can turn to good nutritions to the plants when they decompse; provide shelters for many organisms; and can be good carbon storages. Thus, studying them will give insightinto 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. Trap-based collection: Elevated traps collect small materials like leaves and needles; ground traps collect larger woody debris. 2. Sampling frequency: Elevated traps are sampled biweekly in deciduous forests and every 1-2 months in evergreen forests. Ground traps are sampled once per year. 3. Trap placement: One trap pair per 400m2 is deployed. Traps are placed in 20mx20m or 40mx40m tower plots.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
# Use dim to know the dimenstion of the dataset
dim(Neonics) # row #: 4623, Column #: 30
```

```
## [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...]

```
# # Use summary function and use $ to specify the Effect column
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
```

```
# Use table() to count the effects
table_effect <- table(Neonics$Effect)

# Sort the effects
counts_effect <- sort(table_effect)

print(counts_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 studied effects are population and mortality. The reason why these are mostly studied is because they show the direct impact of the insecticides on ecosystems. High mortality rates indicate toxicity with declines on population.

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...]

```
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
```

```
# Since the last sixth is expressed as other, I assigned the maxsum as 7 to get the sixth species name.
```

Answer: Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble bee, Italian Honeybee. These six species are all bees, which play a role as pollinators helping plant reproduction. These data tells us that neonicotinoids impact on beneficial insects.

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...]

```
# Use class function to know the class of `Conc.1..Author.` column
class(Neonics$Conc.1..Author.)
```

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

```
# Use view function to know the see the `Conc.1..Author.` column
view(Neonics$Conc.1..Author.)
```

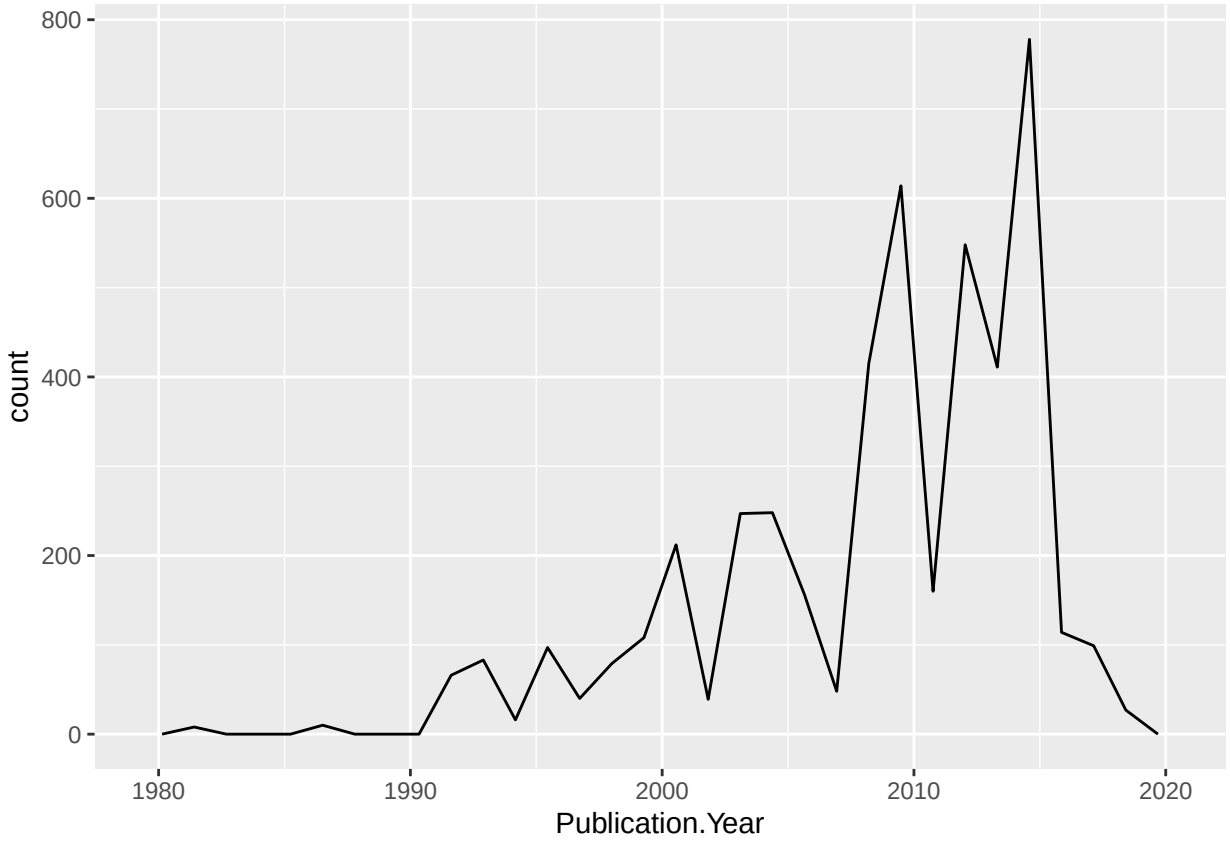
Answer: The class of `Conc.1..Author.` column is a factor. This is because the data has NR (Non-Numeric Entries) and special characters such as the tilde. These makes this data a factor, not numeric.

Explore your data graphically (Neonics)

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

```
# Set x-axis as Publication.Year to generate a plot of the number of studies conducted by publication y
ggplot(Neonics) +
  geom_freqpoly(aes(x=Publication.Year))
```

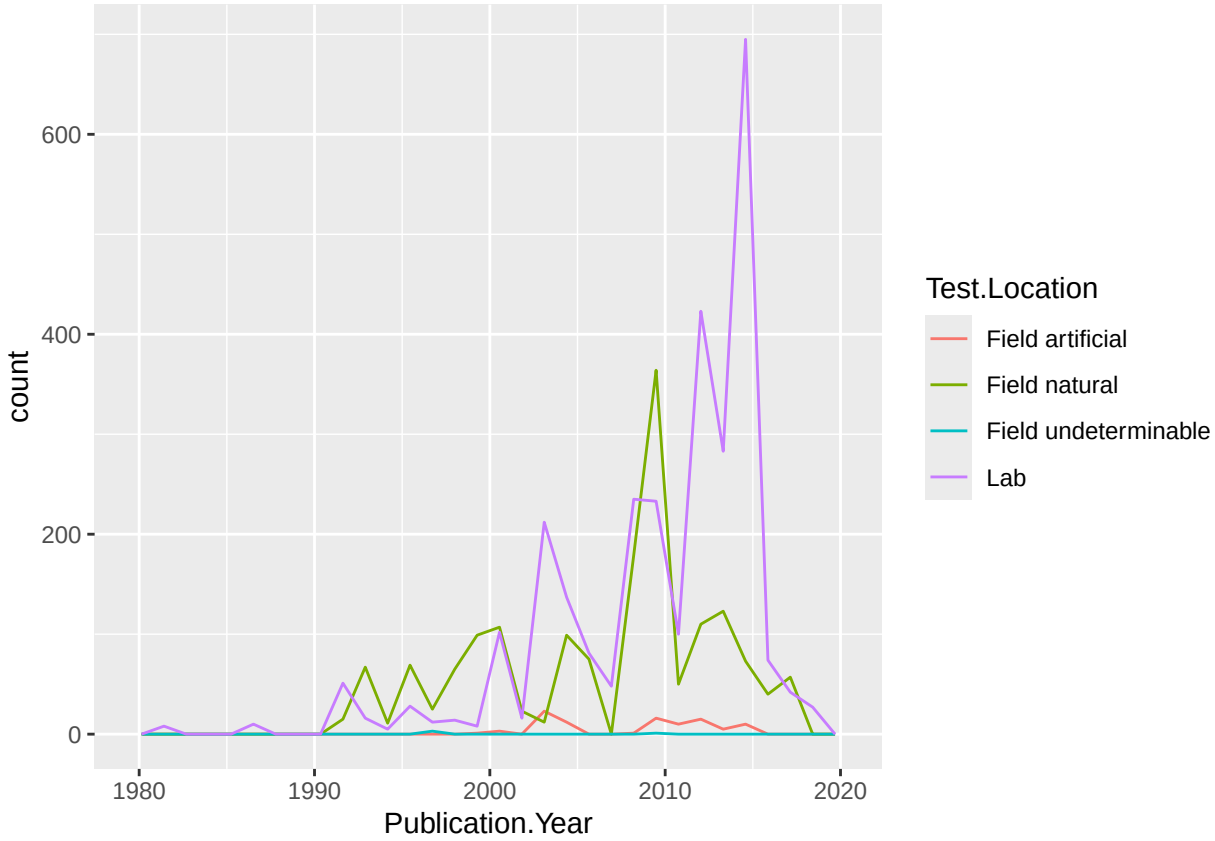
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
# Set x-axis as Publication.Year, and put color function to add Test.Location
ggplot(Neonics) +
  geom_freqpoly(aes(x=Publication.Year, color=Test.Location))
```

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



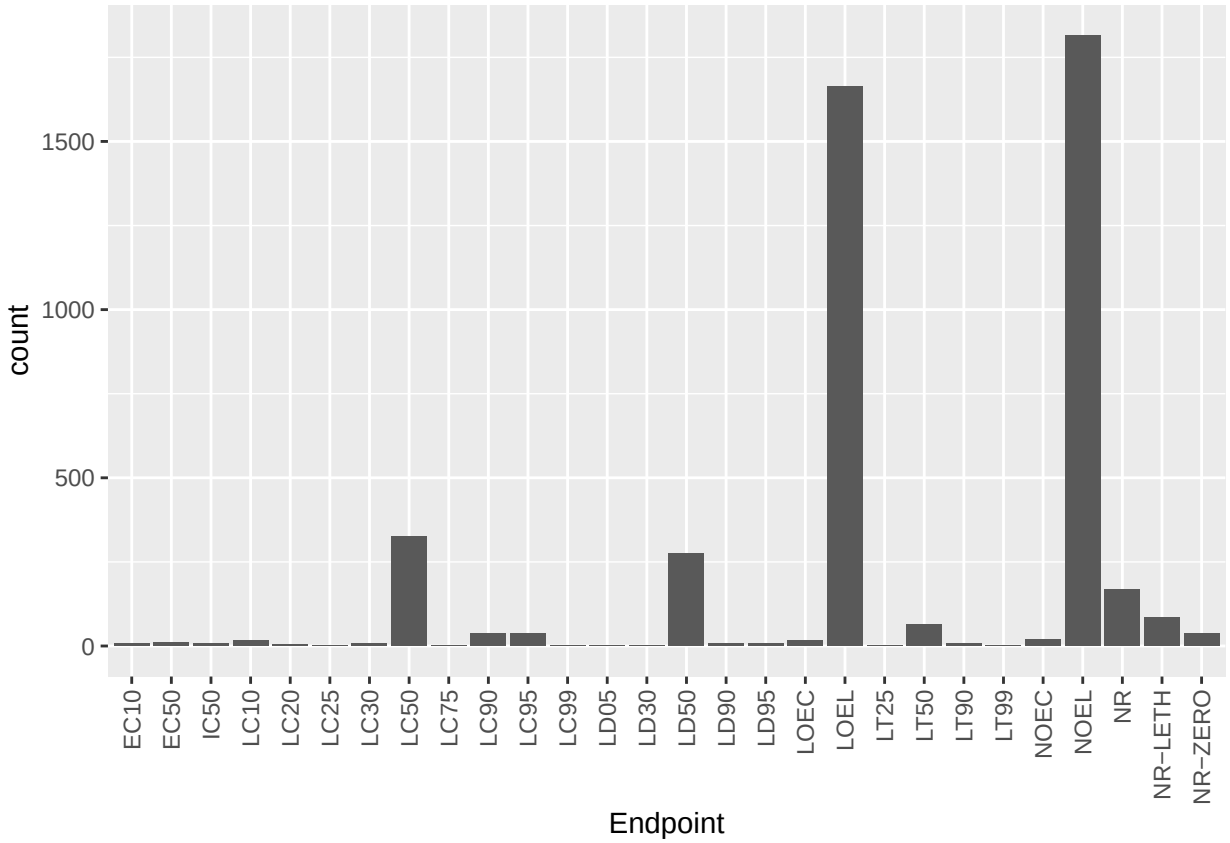
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: Lab and field natural are the most common test locations because their counts are higher than other lines. Lab is getting higher over time; field natural had increased around 2010, and decreased.

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...]

```
# Set x-axis as Endpoint to count them
ggplot(Neonics) +
  geom_bar(aes(x=Endpoint)) +
           theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: (1)NOEL = 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) , (2)LOEL= 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.

```
# Use class function
class(Litter$collectDate) # factor
```

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

```
# Change the class of collectDate as a date
Litter$collectDate <- ymd(Litter$collectDate)
class(Litter$collectDate)
```

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

```
# Changed class
unique(Litter$collectDate) # "2018-08-02" "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
```

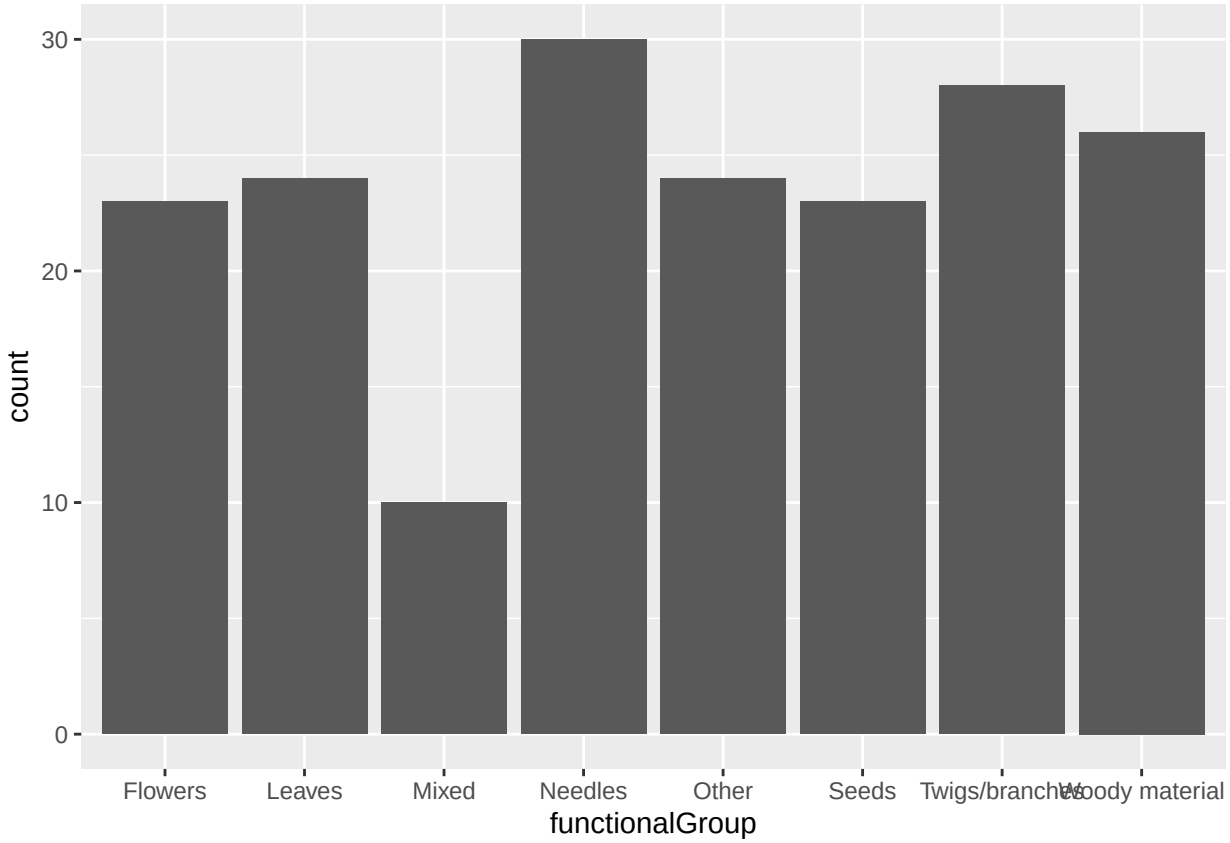
```
summary(Litter$plotID)
```

```
## 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: 12. Data obtained from 'unique' lists all the plot ID and tells the level of the data, while 'summary' tells us all the plot ID and the number of its plot.

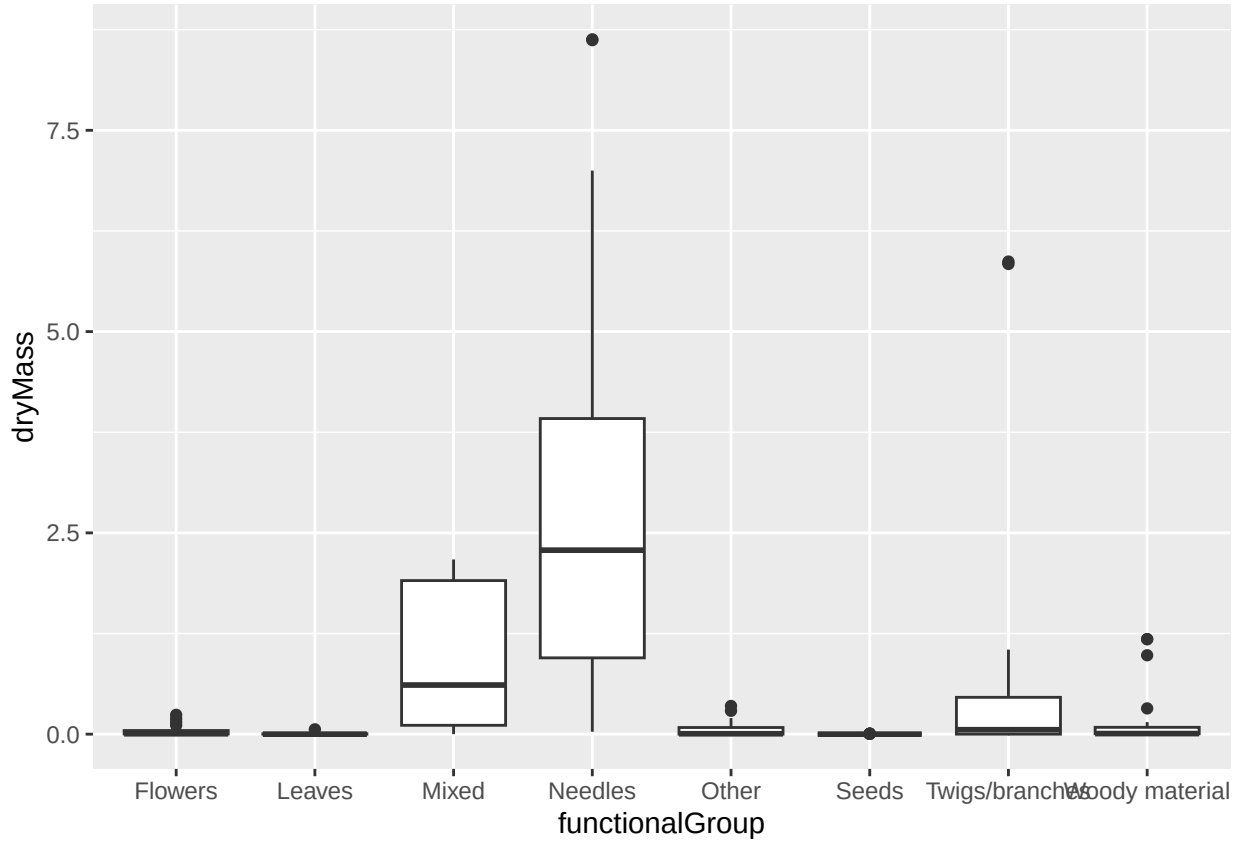
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
# Set x-axis as functionalGroup to count them
ggplot(Litter) +
  geom_bar(aes(x=functionalGroup))
```

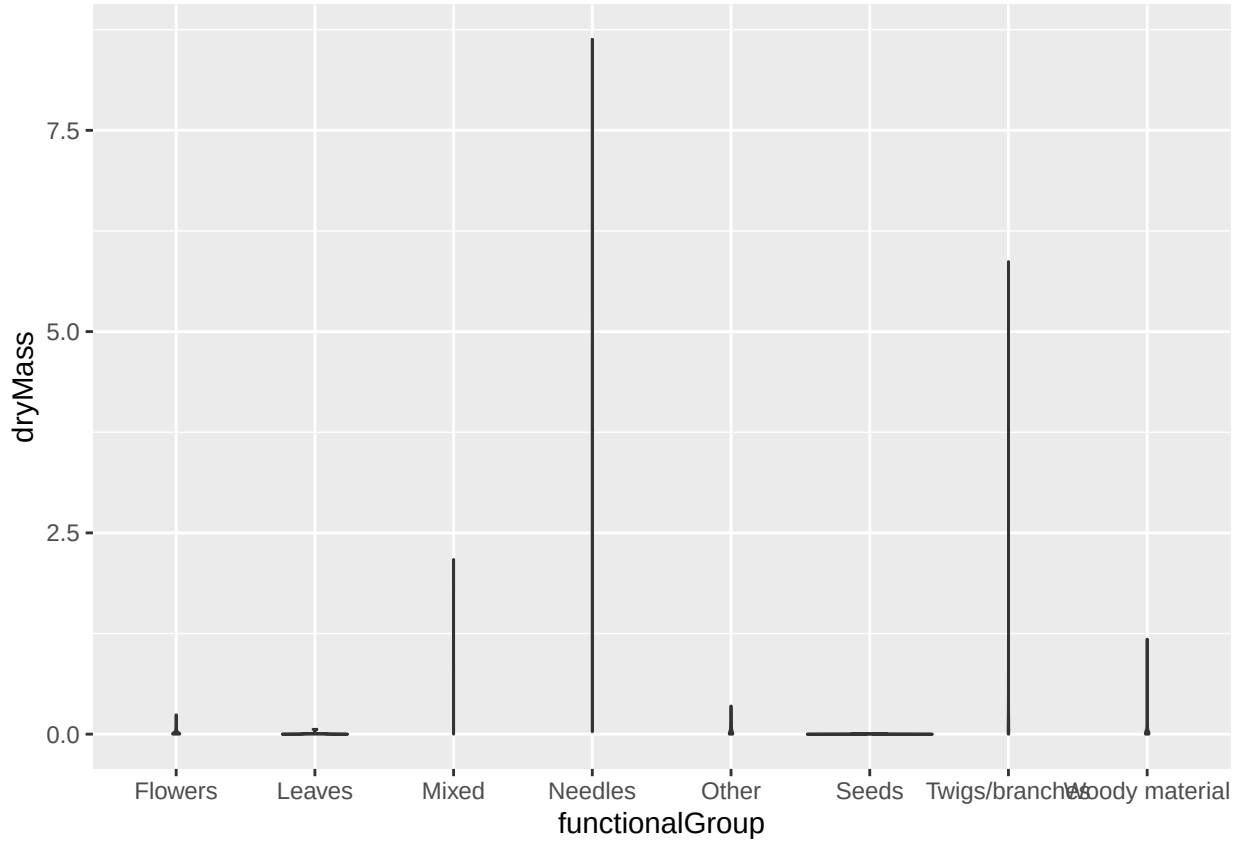


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




```
# Create violin pot
ggplot(Litter) +
  geom_violin(aes(x=functionalGroup, y=dryMass))
```



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

> Answer:Because there is no adequate data to form the violin plot, it's hart to recognize the data distribution. However, through boxplot, I can see some group's median, quantiles, etc.

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

> Answer: Needles. Because it has the highest dryMass among others.