Assignment 3: Data Exploration

Yangsen Zhang

Spring 2026

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. [NEW] 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 Canvas.
8. Initial here to acknowledge that you did not use AI in completing this assignment, except where expressly allowed: Yangsen Zhang

TIP: If your code extends past the page when knit, tidy your code by manually inserting line breaks in your code chunks.

TIP: If your code fails to knit, check: * That no `install.packages()` or `View()` commands exist in your code. * That you are not displaying the entire contents of a large dataframe in your code.

Set up your R session

1. Load necessary packages (tidyverse, here), check your current working directory and import 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: * Use the `here()` package in specifying the paths to your datasets * Include the appropriate subcommand to read in character based columns as factors

```r
# load required packages
library(tidyverse)
library(here)

# check current working directory
getwd()
```

```
## [1] "/home/guest/EDE_Spring2026-yz"
```

```r
# import 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. (AI is allowed here, but put answers in your own words.)

   Answer: We might be interested in the ecotoxicology of neonicotinoids on insects because although these pesticides are effective against pests, they may also harm other non-target insect species, such as bees. Understanding this can help us understand their role in ecosystems and others.

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. (AI is allowed here, but put answers in your own words.)

   Answer:Litter and woody debris that falls to the ground in forests is crucial to carbon and nutrient cycle and soil health, because when leaves, branches and other plant substances fall and decompose, they will release the necessary nutrients back to the soil to support diversified decomposer communities. They also help to maintain soil moisture, prevent erosion, and play a role in the carbon cycle. Studying these processes helps us understand how forest ecosystems work and respond to environmental changes. These indicators can help us assess forest's overall function and health in a changing world.

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.Sampling is conducted using elevated and ground traps 2.Trap placement within plots may be either targeted or randomized, depending on the vegetation 3.The sampling schedule depends on the trap type and the ecosystem. Ground traps are sampled only once per year. Elevated litter traps are sampled more frequently, but the rate varies

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

This dataset contains 4623 observations and 30 variables.

6. Using the `summary` function on the "Effect" column, determine the most common effects that are studied. [Tip: The `sort()` command is useful for listing the values in order of magnitude...]

```
# What class is the Effect column now?
class(Neonics$Effect)
```

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

```
# Summary 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
```

```
# Sort effects by frequency
sort(summary(Neonics$Effect), decreasing = TRUE)
```

```
##       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
```

Based on the summary of the Effect column, Population is the most commonly studied effect, followed by mortality and behavior.

Question: Which two effects stand out as the most studied? Can you guess why these effects might specifically be of interest? > Answer: The two most commonly studied effects are population and mortality. Population are of particular interest because changes in a population are often the most direct response to the environmental changes, such as pollution, climate change and so on. Studying mortality may help to find the reasons of population decline, it is useful to protect species.

7. Using the `summary` function, determine the six most commonly studied species in the dataset (common name).[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
```

The six most commonly studied species are Honey bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee.

Question: What do these species have in common? Why might they be of interest over other insects? > Answer:These species are primarily pollinators.They are of interest might because neonicotinoids can negatively affect pollinator health, with implications for ecosystem function and agriculture.

8. The `Conc.1..Author` column, which lists the concentration of the neonicitoid dose, should include numeric values. 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...]

```
# check the class of `Conc.1..Author.`
class(Neonics$Conc.1..Author)
```

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

```
# View the first few values
head(Neonics$Conc.1..Author)
```

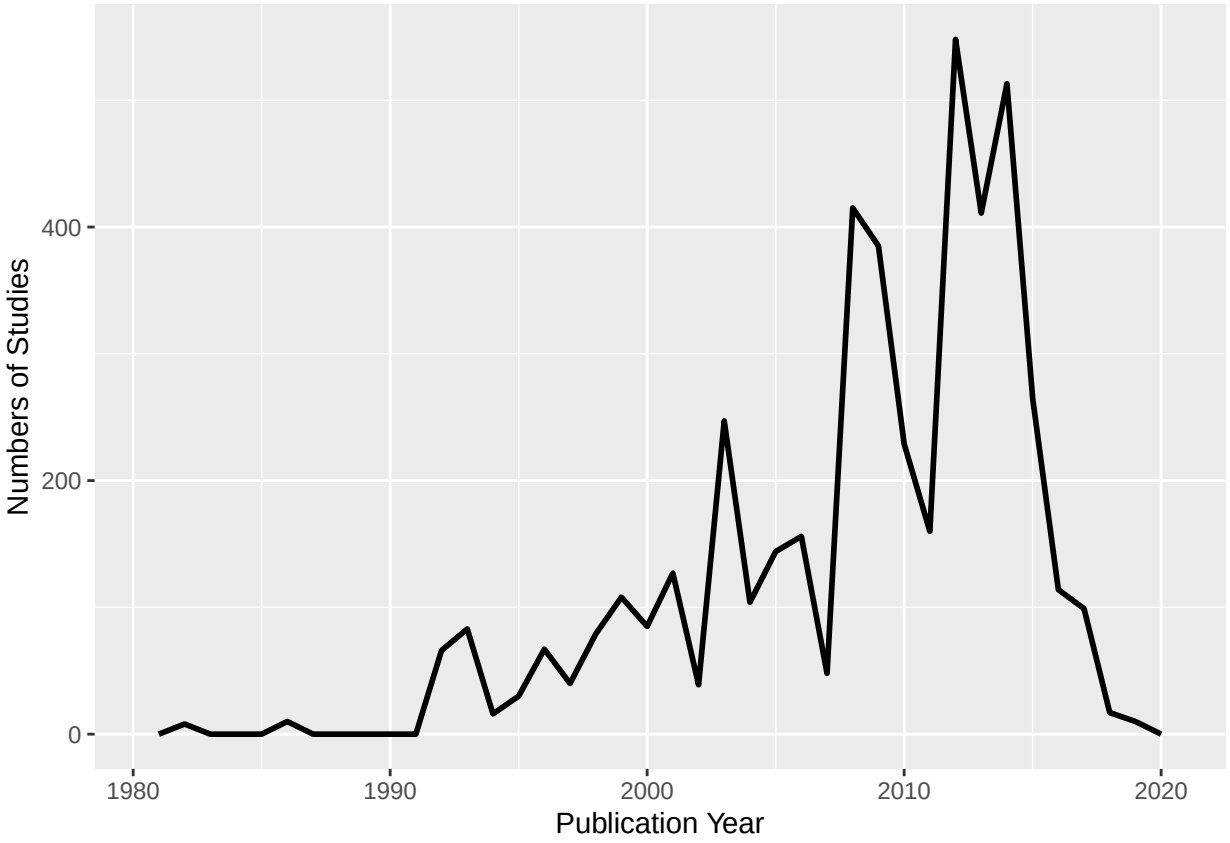
```
## [1] 27.2 19.7 47   25   13   268
## 1006 Levels: <0.0004 <0.025 <0.088 <0.5 <1.5 <10/ <2.5/ <4.00 <5.00 ... NR/
```

Answer:This column is stored as a factor rather than a numeric variable, That's because this column contains non-numeric values such as "<0.0004" "NR".

Explore your data graphically (Neonics)

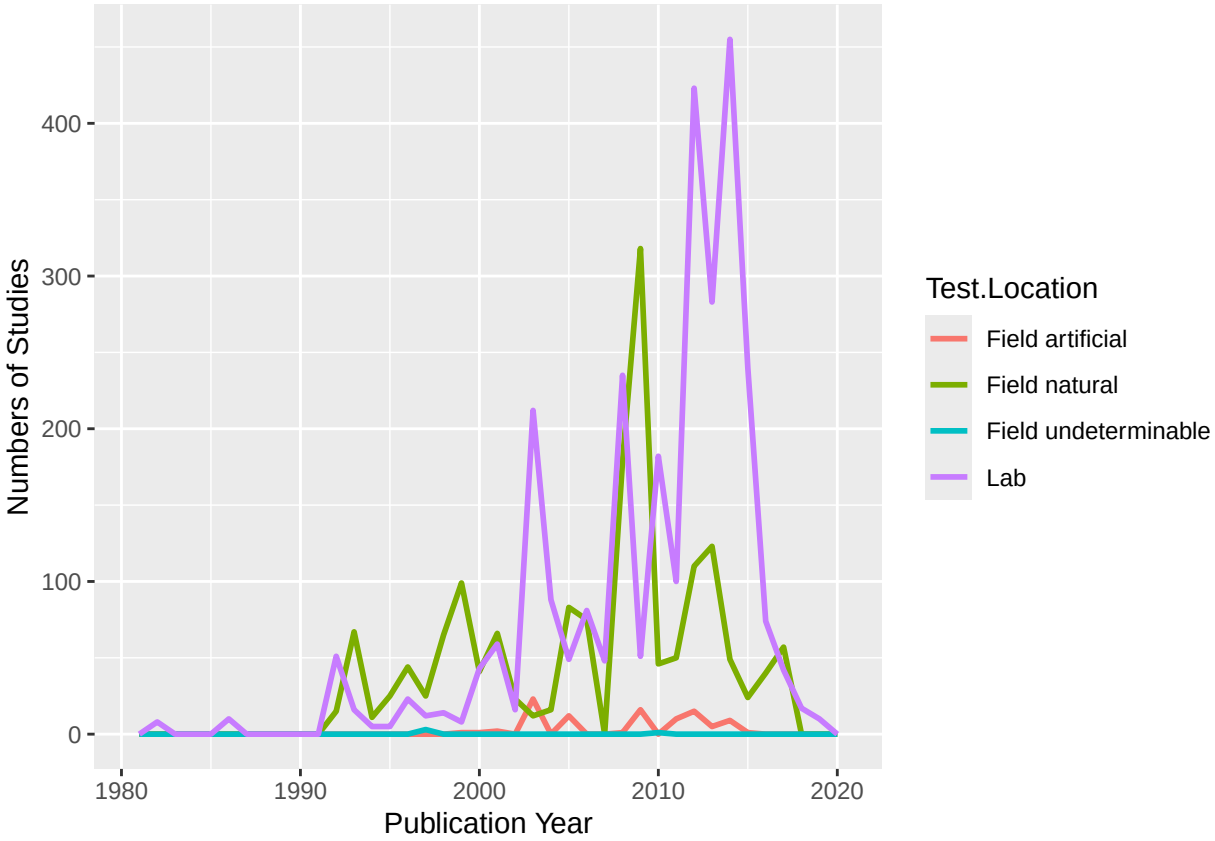
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
# Plot number of studies by publication year
ggplot(Neonics,aes(x = Publication.Year)) +
  geom_freqpoly(binwidth =1, color = "black", linewidth = 1) +
  labs( x = "Publication Year", y = "Numbers of Studies")
```



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

```
# Plot number of studies by publication year
ggplot(Neonics, aes(x = Publication.Year, color = Test.Location)) +
  geom_freqpoly(binwidth =1, linewidth = 1) +
  labs( x = "Publication Year", y = "Numbers of Studies")
```

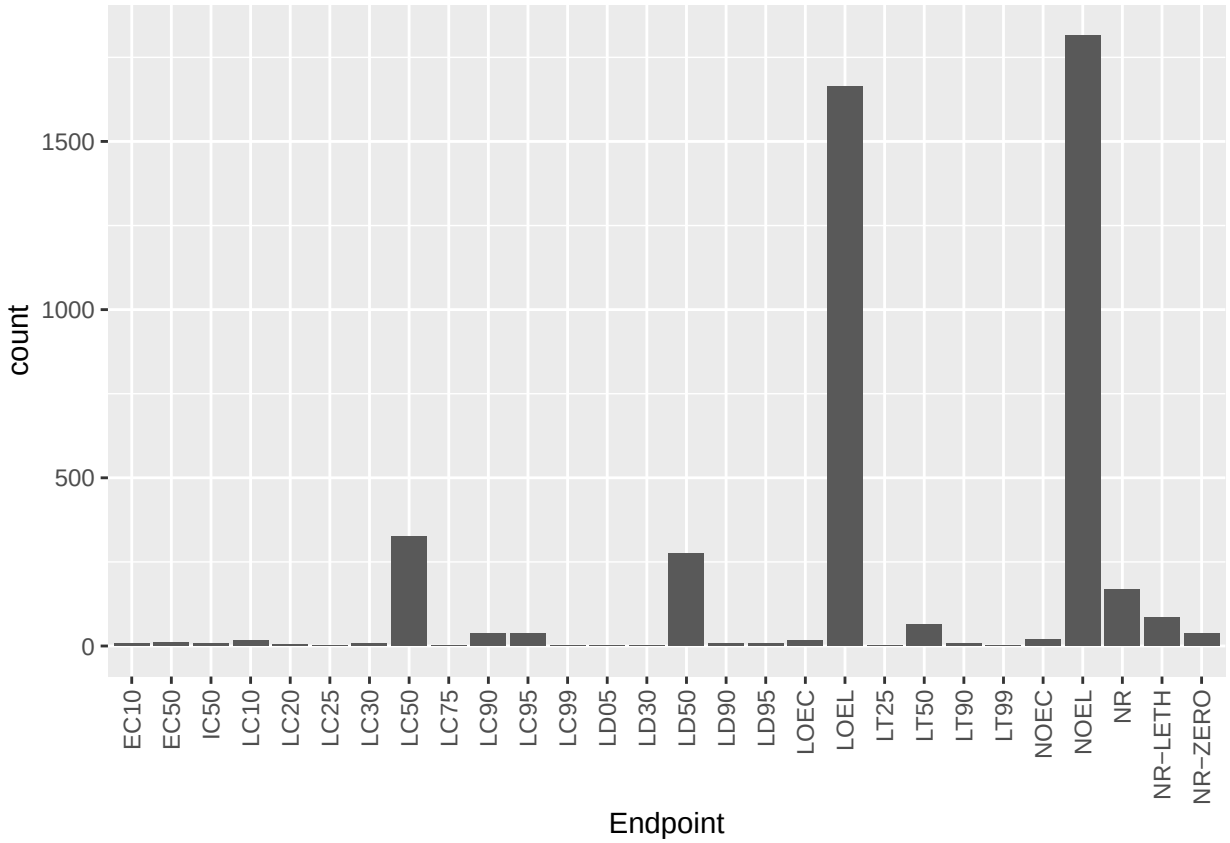


Interpret this graph. What are the most common test locations, and do they differ over time? > Answer:The most common test location is Lab.The relative importance of test locations changes over time, with laboratory-based studies becoming increasingly dominant in later years

11. Create a bar graph of Endpoint counts.

[**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...]

```
# Create a bar graph of Endpoint counts.
ggplot(Neonics,aes(x = Endpoint)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



What are the two most common end points, and how are they defined? Consult the ECOTOX_CodeAppendix (p.721) for more information. > Answer:The two most common endpoints in the dataset are NOEL and LOEL. No-observable-effect-level(NOEL) is the highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test. Lowest-observable-effect-level(LOEL) is the lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls.

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.

```
# check the class
class(Litter$collectDate)
```

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

```
# view the first few values of collectDate
head(Litter$collectDate)
```

```
## [1] 2018-08-02 2018-08-02 2018-08-02 2018-08-02 2018-08-02 2018-08-02
## Levels: 2018-08-02 2018-08-30
```

```
# load the lubridate package
library(lubridate)
# convert collectDate from factor to Date format
Litter$collectDate <- ymd(Litter$collectDate)
# comfirm the class
class(Litter$collectDate)
```

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

```
# identify which dates litter was sampled in August 2018
unique(Litter$collectDate)
```

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

13. Using the `unique` function, list the different `plotIDs` sampled at Niwot Ridge.

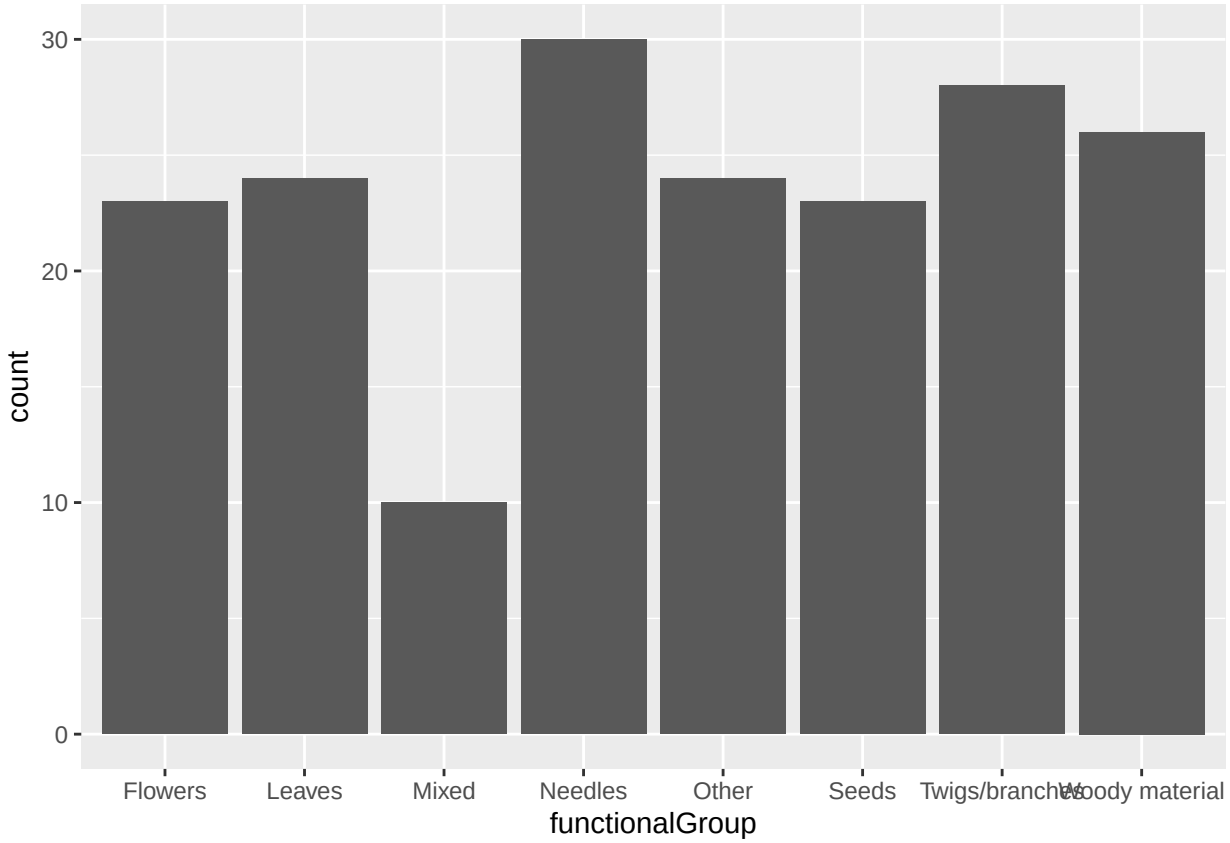
```
# list the different `plotIDs` sampled at Niwot Ridge
unique(Litter$plotID)
```

```
##  [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
```

How is the information obtained from `unique` different from that obtained from `summary`? > Answer:Unique function returns to a list of all distinct plotIDs in the dataset, while summary function summarizes the data by showing the frequency of observations for each plotID.

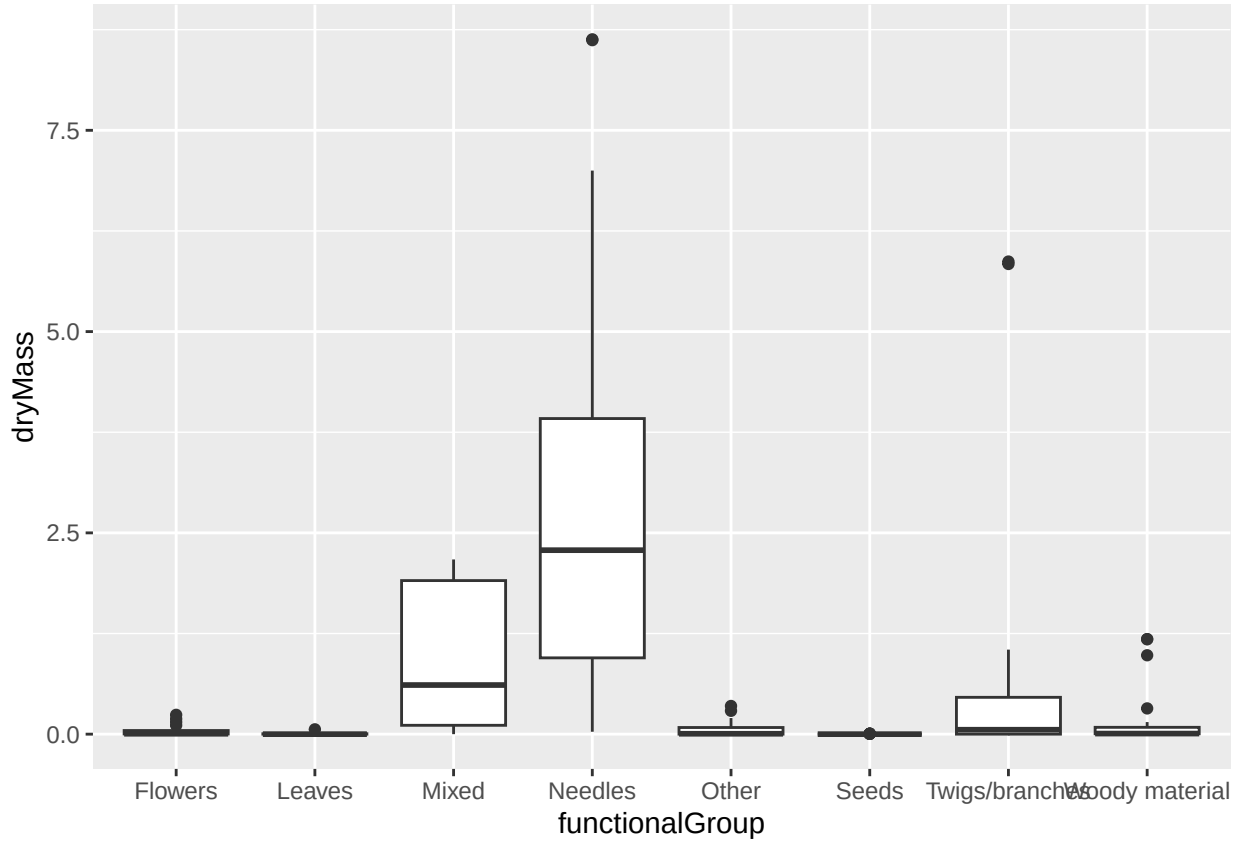
14. Create a bar graph of `functionalGroup` counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
# create a bar graph of `functionalGroup` counts
ggplot(Litter,aes(x = functionalGroup)) +
  geom_bar()
```

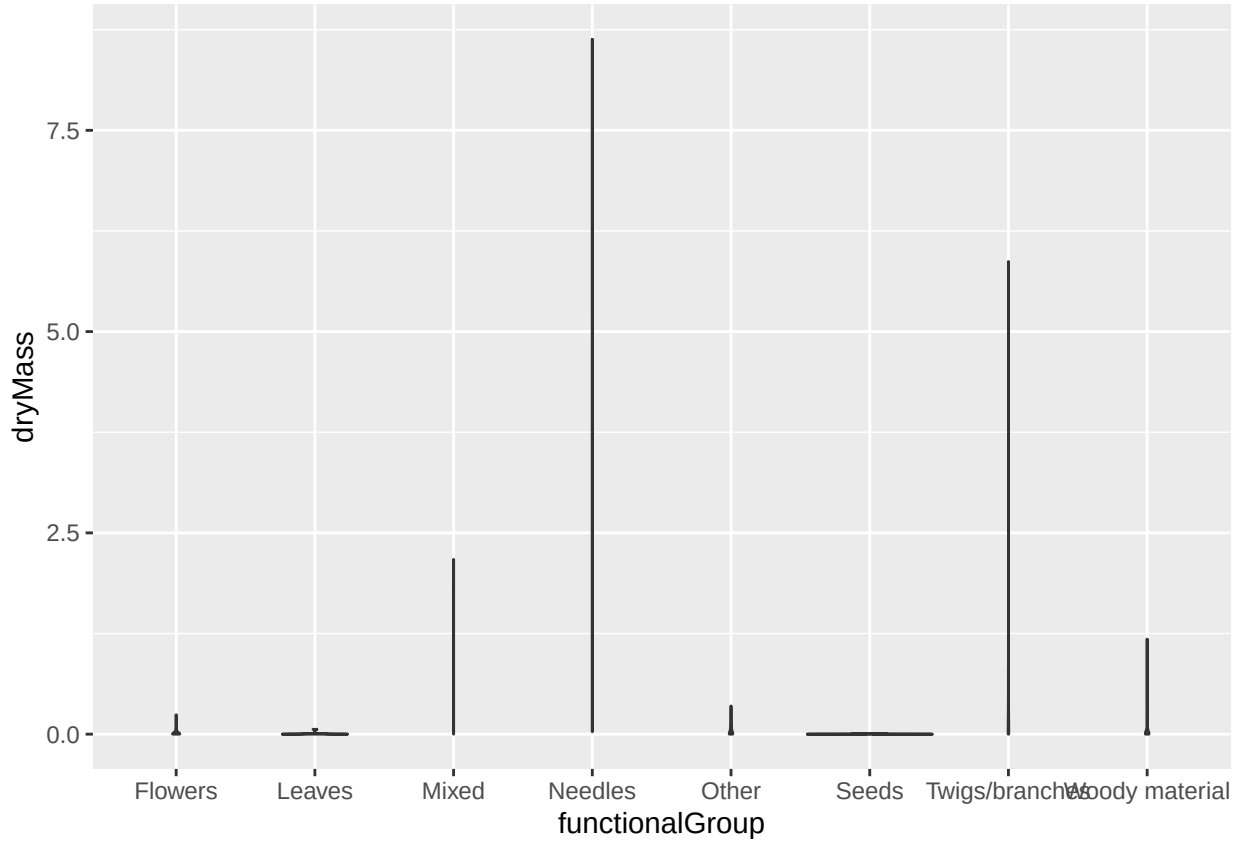


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of `dryMass` by `functionalGroup`.

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




```
# create a violin plot
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_violin()
```



Why is the boxplot a more effective visualization option than the violin plot in this case? > Answer:The boxplot shows central tendency, spread, and critical outliers for data that is often skewed and contains many zero values, without the potential for misinterpretation that a violin plot's smoothing might cause.

What type(s) of litter tend to have the highest biomass at these sites? > Answer:Needles tend to have the highest biomass at these site.