Assignment 3: Data Exploration

He Gao

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: He

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
library(tidyverse)
library(here)
here()
```

```
## [1] "C:/Users/ /EDE_Spring2026"
```

```r
Neonics <- read.csv(
  here("Data", "Raw", "ECOTOX_Neonicotinoids_Insects_raw.csv"),
  stringsAsFactors = TRUE
)
Litter <- read.csv(
  here("Data", "Raw", "NEON_NIWO_Litter_massdata_2018-08_raw.csv"),
  stringsAsFactors = TRUE
)
dim(Neonics)
```

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

```r
dim(Litter)
```

```
## [1] 188  19
```

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: It's important to study the effects of neonicotinoids on insects for a few key reasons. First, insects are crucial for ecosystems—they pollinate crops, decomposition, and serving as a foundation of food webs. Second, these insecticides don't just kill pest insects; they can also harm beneficial ones like bees and other pollinators. Finally, understanding these impacts helps scientists and policymakers assess the risks and create regulations that protect our environment based on solid evidence.

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:Forest litter and woody debris play an important role in how forest ecosystem.First, they help store carbon. Leaves, branches, and dead wood contain large amounts of carbon, and keeping this material in the forest helps reduce the amount of carbon released into the atmosphere.Second, they support nutrient cycling. As plant material slowly decomposes, nutrients such as nitrogen and phosphorus are released into the soil. These nutrients can then be reused by plants for future growth.Third, litter and woody debris contribute to soil formation. Decomposed organic matter improves soil structure and fertility, which helps plants grow better.Because of these functions, studying forest litter and woody debris helps scientists better understand forest productivity, carbon movement, and how forests respond to climate change and other environmental pressures over time.

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.Use of two complementary trap types NEON uses elevated litter traps to collect leaves and small plant material, and ground traps to collect longer fine woody debris. This design ensures that different types and sizes of falling plant material are sampled effectively. 2.Standardized plot-based sampling design Sampling is conducted within fixed plots at NEON sites with woody vegetation. Trap pairs are placed following standardized rules, with either random or targeted placement depending on vegetation cover, which allows data to be comparable across sites. 3.Seasonally adjusted sampling frequency Litter traps are collected more frequently during periods of high litterfall, such as autumn in deciduous forests, while ground traps for woody debris are collected less frequently, typically once per year. This approach reflects natural differences in how plant material falls over time.

4. NEON litterfall sampling methods (three key points)

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

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...]

```
effect_summary <- summary(Neonics$Effect)
sort(effect_summary, 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
```

Question: Which two effects stand out as the most studied? Can you guess why these effects might specifically be of interest? > Answer: The two effects that are studied most often are mortality and population effects. Population effects are also important, since they show how insect populations may be affected over time, rather than only focusing on immediate deaths. Mortality is commonly studied because it is easy to observe and provides a direct indication of toxicity. Together, these two effects help researchers understand both the short-term and longer-term impacts of neonicotinoids.

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…]

```
common_col <- names(Neonics)[grepl("common", names(Neonics), ignore.case = TRUE)][1]
summary(Neonics[[common_col]], maxsum = 6)
```

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

```
# Note: The output was NULL when I use  `summary()` function, in particular the `maxsum` argument ,so I
```

Question: What do these species have in common? Why might they be of interest over other insects? > Answer:These species have in common that they are all pollinating insects. Bees and parasitic wasps play an important role in pollination and ecosystem functioning. They are closely linked to plant reproduction and food production, and changes in their survival or population can have wide ecological and agricultural impacts. Compared to many other insects, these species are also more sensitive to pesticide exposure, which makes them important indicators when studying the effects of neonicotinoids.

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…]

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

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

```
head(Neonics$Conc.1..Author, 20)
```

```
##  [1] 27.2   19.7   47     25     13     268    170    28     48     40     83     900
## [13] 15.3   20.4   5      5      NR     ~10    65.56  635.4
## 1006 Levels: ~10 ~30/ ~40/ ~41 <0.0004 <0.025 <0.088 <0.5 <1.5 <10/ ... NR/
```

> Answer:This column is not numeric because it contains non-numeric entries like "NR" and values with inequality signs (e.g., "<10"), so R treats it as a categorical (factor) variable instead of a numeric one.

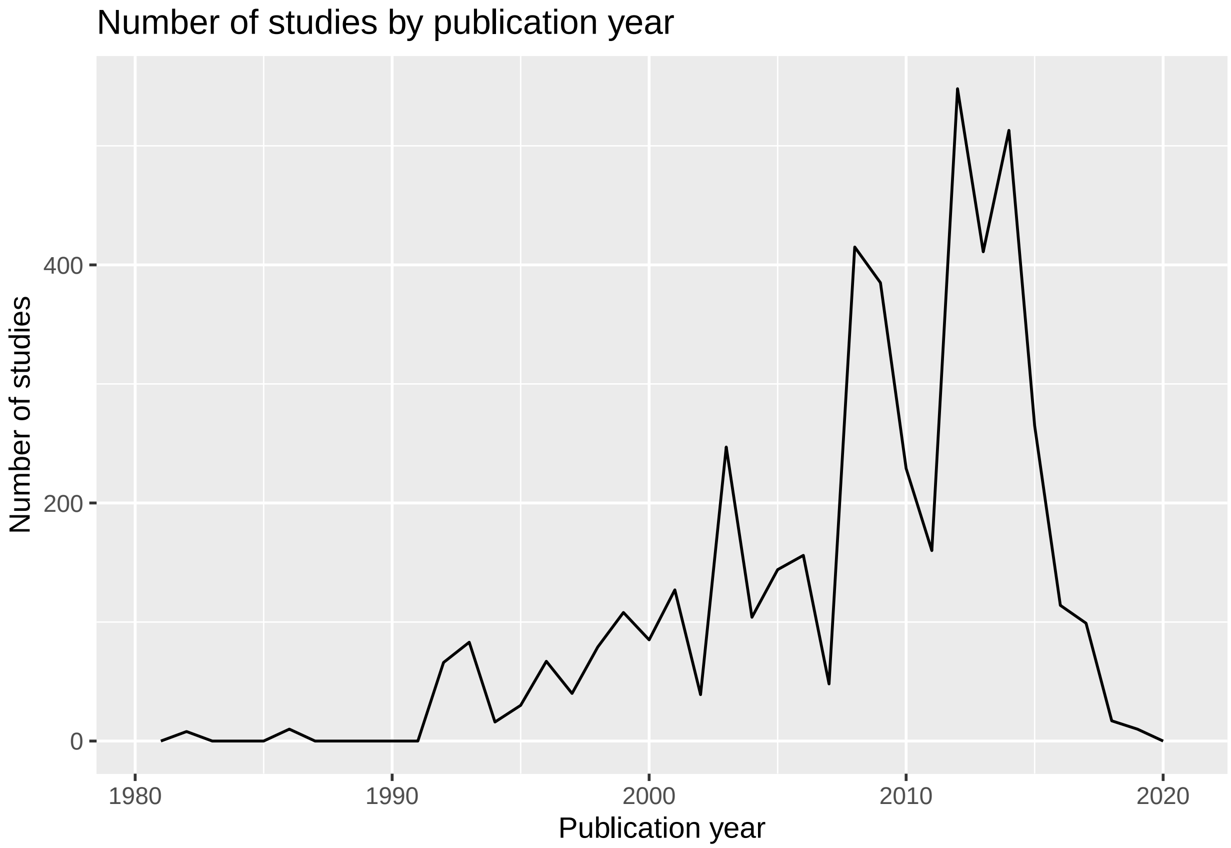
Explore your data graphically (Neonics)

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

```
year_col <- names(Neonics)[grepl("Year", names(Neonics), ignore.case = TRUE)][1]
year_col
```

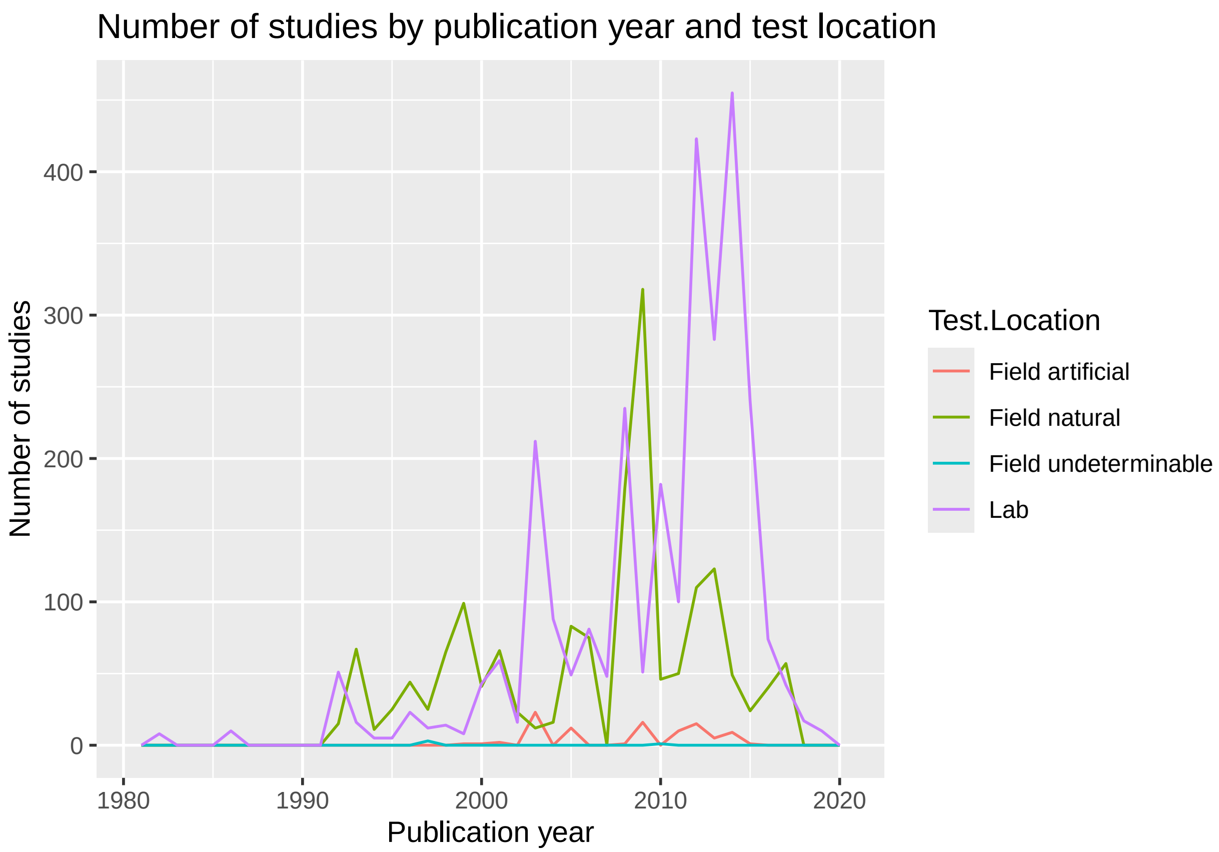
```
## [1] "Publication.Year"
```

```r
ggplot(Neonics, aes(x = .data[[year_col]])) +
  geom_freqpoly(binwidth = 1) +
  labs(
    x = "Publication year",
    y = "Number of studies",
    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
ggplot(Neonics, aes(x = Publication.Year, color = Test.Location)) +
  geom_freqpoly(binwidth = 1) +
  labs(
    x = "Publication year",
    y = "Number of studies",
    title = "Number of studies by publication year and test location"
  )
```

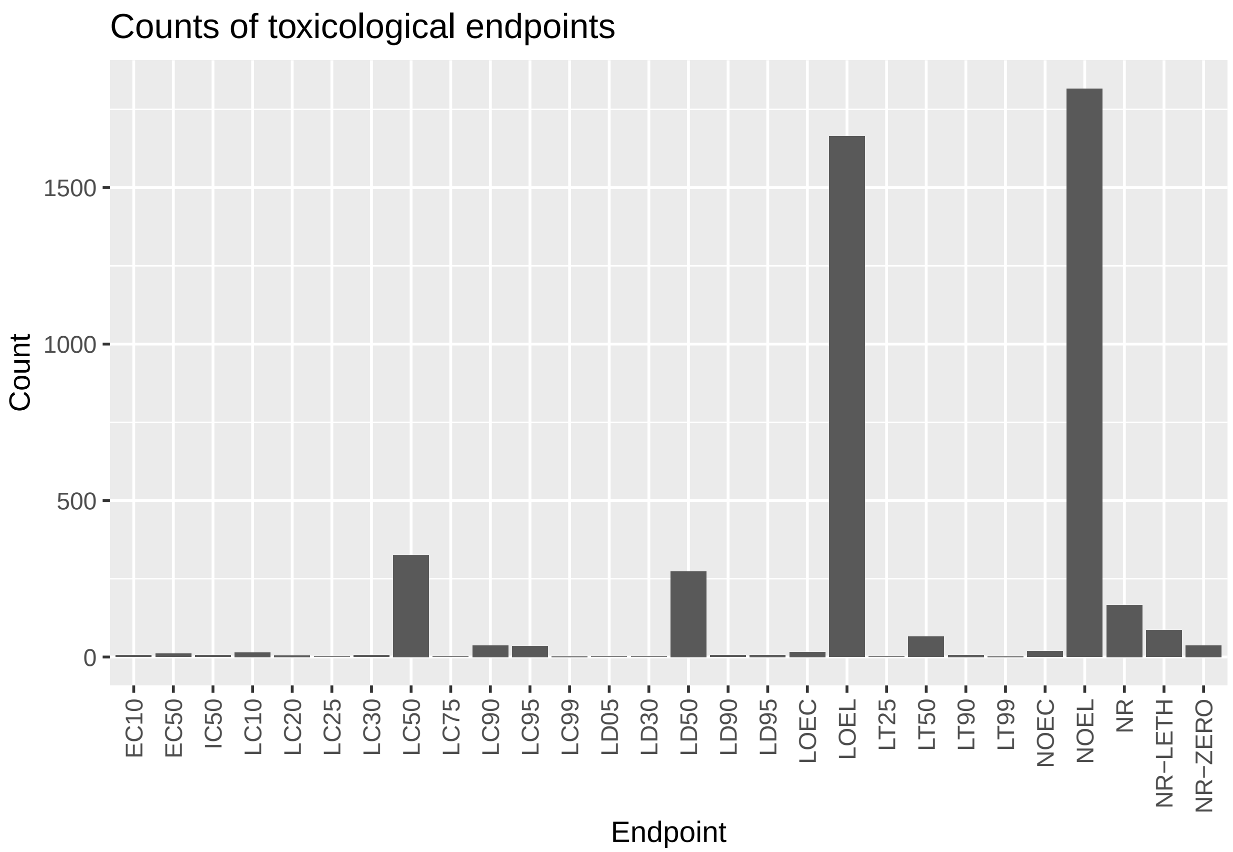


Interpret this graph. What are the most common test locations, and do they differ over time? > Answer: The lab is clearly the most common test location. Field natural studies are also fairly common but much less frequent than lab studies, while the other field categories appear occasionally. In the late 2000s and early 2010s, the number of studies increases but the overall pattern stays similar, with lab studies remaining the most common.

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…]

```
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar() +
  theme(
    axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)
  ) +
  labs(
    x = "Endpoint",
    y = "Count",
    title = "Counts of toxicological endpoints"
  )
```



What are the two most common end points, and how are they defined? Consult the ECOTOX_CodeAppendix (p.721) for more information? The two most common endpoints are NOEL and LOEL. NOEL means the highest dose (or concentration) where the organism's response is not statistically different from the control group (i.e., "no clear effect compared with normal conditions").LOEL means the lowest dose (or concentration) where the organism shows an effect that is statistically different from the control group (i.e., this is the first dose where an effect becomes detectable).Together, they help describe a threshold: when effects begin and when there are still no effects.

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"
```

```
Litter$collectDate <- as.Date(Litter$collectDate)
class(Litter$collectDate)
```

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

```
aug_dates <- unique(Litter$collectDate)
aug_dates
```

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

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

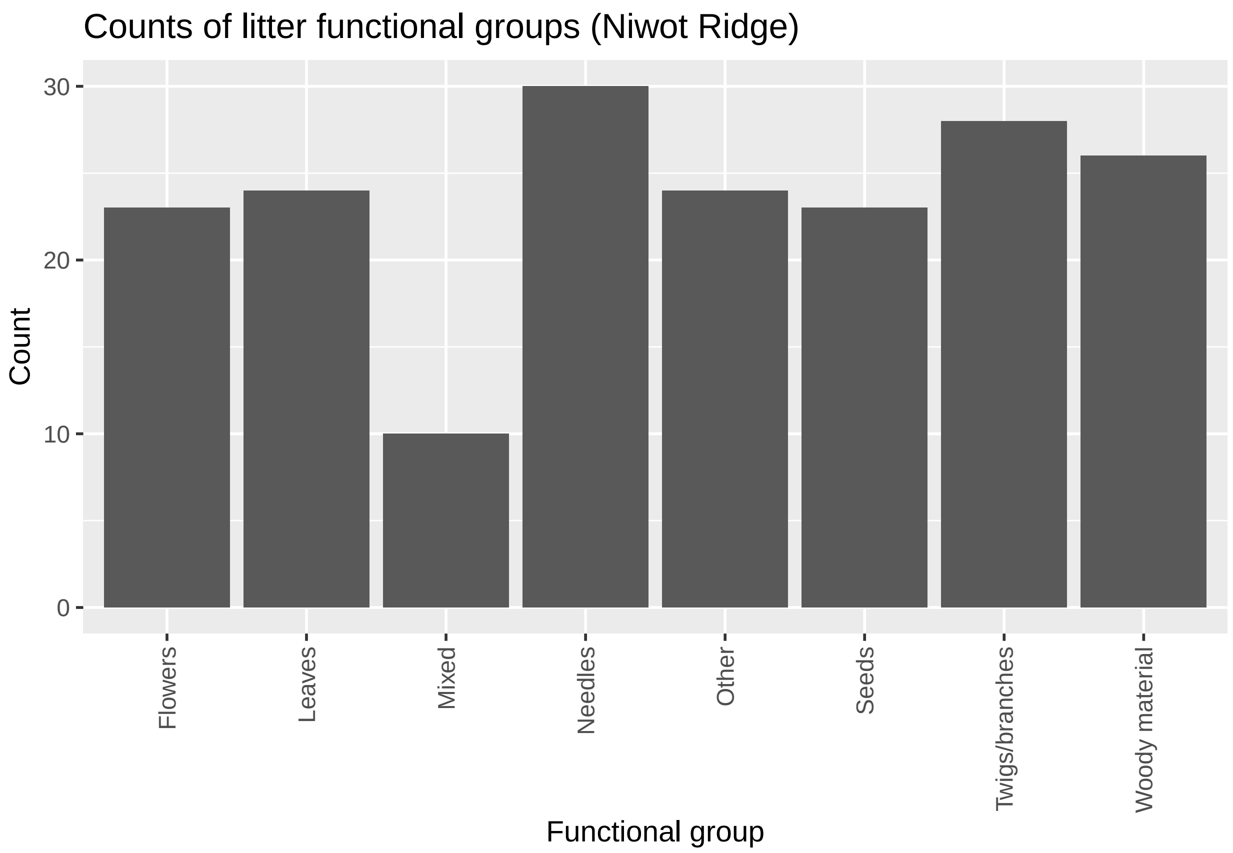
```
unique(Litter$plotID)
```

```
##  [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() shows the different values that appear in a column but it does not tell how many times each value appears. In contrast, summary() gives a count or distribution of the values.

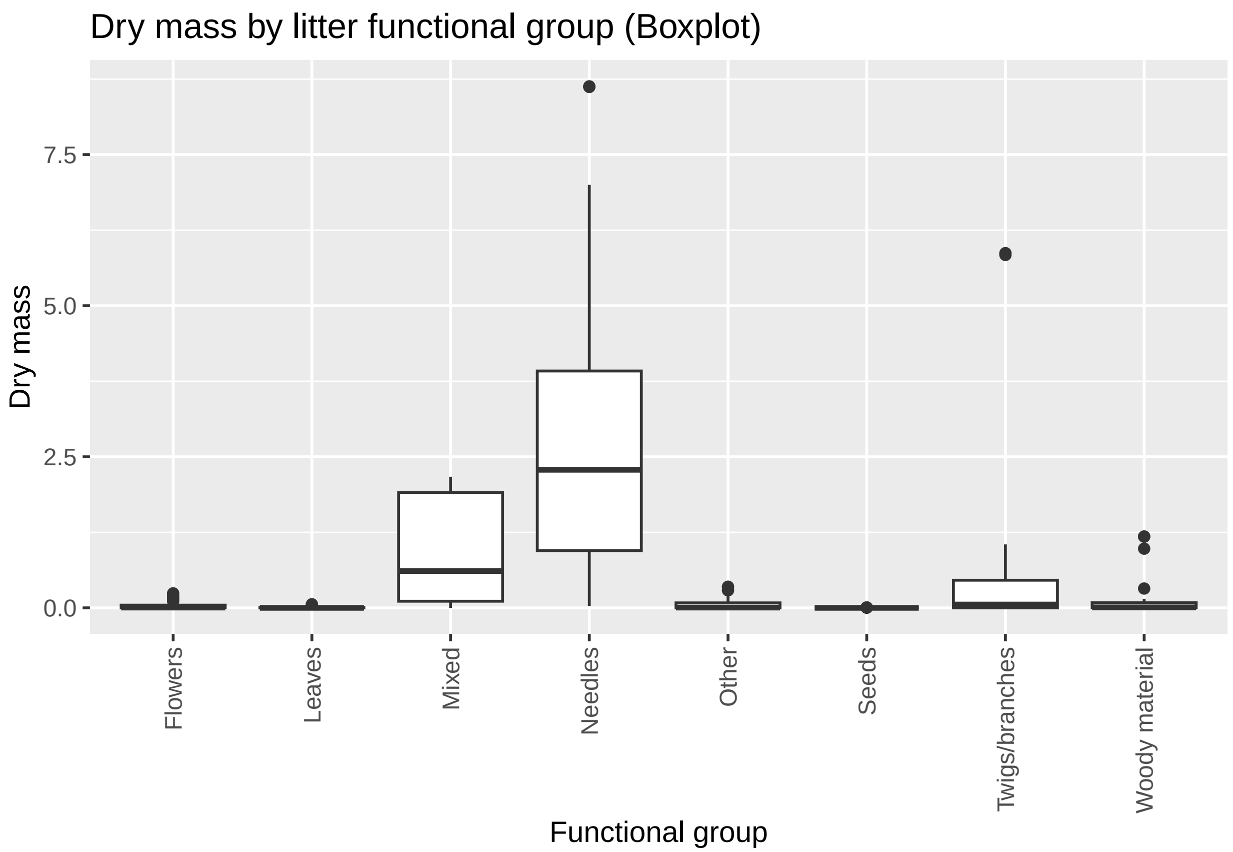
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar() +
  labs(
    x = "Functional group",
    y = "Count",
    title = "Counts of litter functional groups (Niwot Ridge)"
  ) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```

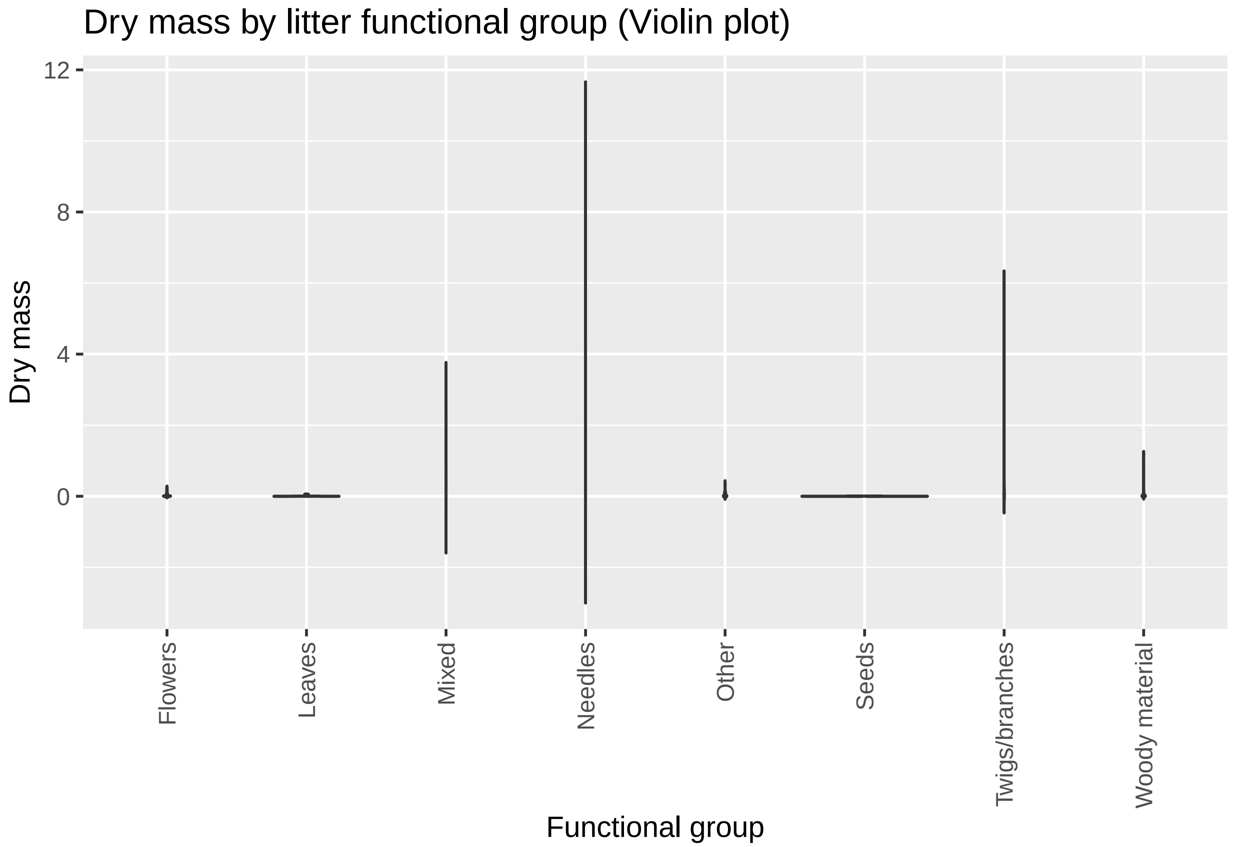


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of `dryMass` by `functionalGroup`.

```
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_boxplot() +
  labs(
    x = "Functional group",
    y = "Dry mass",
    title = "Dry mass by litter functional group (Boxplot)"
  ) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



```r
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_violin(trim = FALSE) +
  labs(
    x = "Functional group",
    y = "Dry mass",
    title = "Dry mass by litter functional group (Violin plot)"
  ) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



Why is the boxplot a more effective visualization option than the violin plot in this case? > Answer: The boxplot is more effective here because it is straightforward: it clearly shows the median, the typical spread of values, and any outliers for each functional group, making groups easy to compare. The violin plot is meant to highlight the shape of the distribution, but when the sample size is small it may look “noisy,” and the differences between groups are harder to read.

What type(s) of litter tend to have the highest biomass at these sites? > Answer: At these sites, the highest biomass tends to come from the woody litter types, especially twigs/branches and woody material. These pieces are physically heavier than softer litter like leaves, needles, or flowers, so they contribute more to total dry mass.