Assignment 3: Data Exploration

Haonan Pei

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: H.P

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 necessary 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
```

```r
library(here)
```

```
## here() starts at /home/guest/EDE_Spring2026
```

```r
#Check current working directory
getwd()
```

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

```r
#Import two dataset
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: The reason why we study the ecotoxicology of neonicotinoids is because these chemicals can harm insects that are vital for ecosystem and agriculture, even at low concentrations. The study on this allow scientists and policymakers to understand the risks of widespread use neonicotinoids (high environmental exposure) and contribute to protect biodiversity and ecosystem.

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: Studying forest litter and woody debris can help scientists to understand nutrient cycling, carbon storage, habitat structure, water dynamics, and ecosystem change related to the forests. It's a foundational information of how forests respond to natural and human-driven influences. We are able to regulate soil, water, ecosystem, and even climate based on the information colletced from the forest litter and woody debris.

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. NEON has two different types of traps to collect different materials: elevated litter traps to collect leaves, needles, and small twigs, and ground traps to collect fine woody debris. 2. Sampling occurs within standardized tower plots using a fixed spatial design that includes a set of elevated litter trap and a set of ground trap within a 20mx20m plot. 3. Ground traps are sampled once per year, while elevated litter traps are sampled more frequently (up to once every two weeks).

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
# Number of rows
nrow(Neonics)
```

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

```
# Number of columns
ncol(Neonics)
```

```
## [1] 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...]

```
# Most common effects
Neonic_effect <- summary(Neonics$Effect)
sort(Neonic_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
```

Question: Which two effects stand out as the most studied? Can you guess why these effects might specifically be of interest? Answer: The effects on the population and mortality stand out as the most studied effects. These effects are specifically interested because they are directly related to whether a population of insects can survive from the environmental exposure of neonicotinoids. The population count and the mortality are measurable indicators that can reflect the health situation of a population of insects over time.

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...]

```
# Six most commonly studied species
Neonic_species <- summary(Neonics$Species.Common.Name, maxsum = 7)
Neonic_species
```

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

Question: What do these species have in common? Why might they be of interest over other insects? Answer: The species that most commonly studied belongs to the family of bee. They are most interested because they are playing critical role in pollinating plants, which makes them ecologically important. Declines in bee health can directly affect crop pollination and food production, making them to be the high priority in ecotoxicological research compared to other insects.

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.)
```

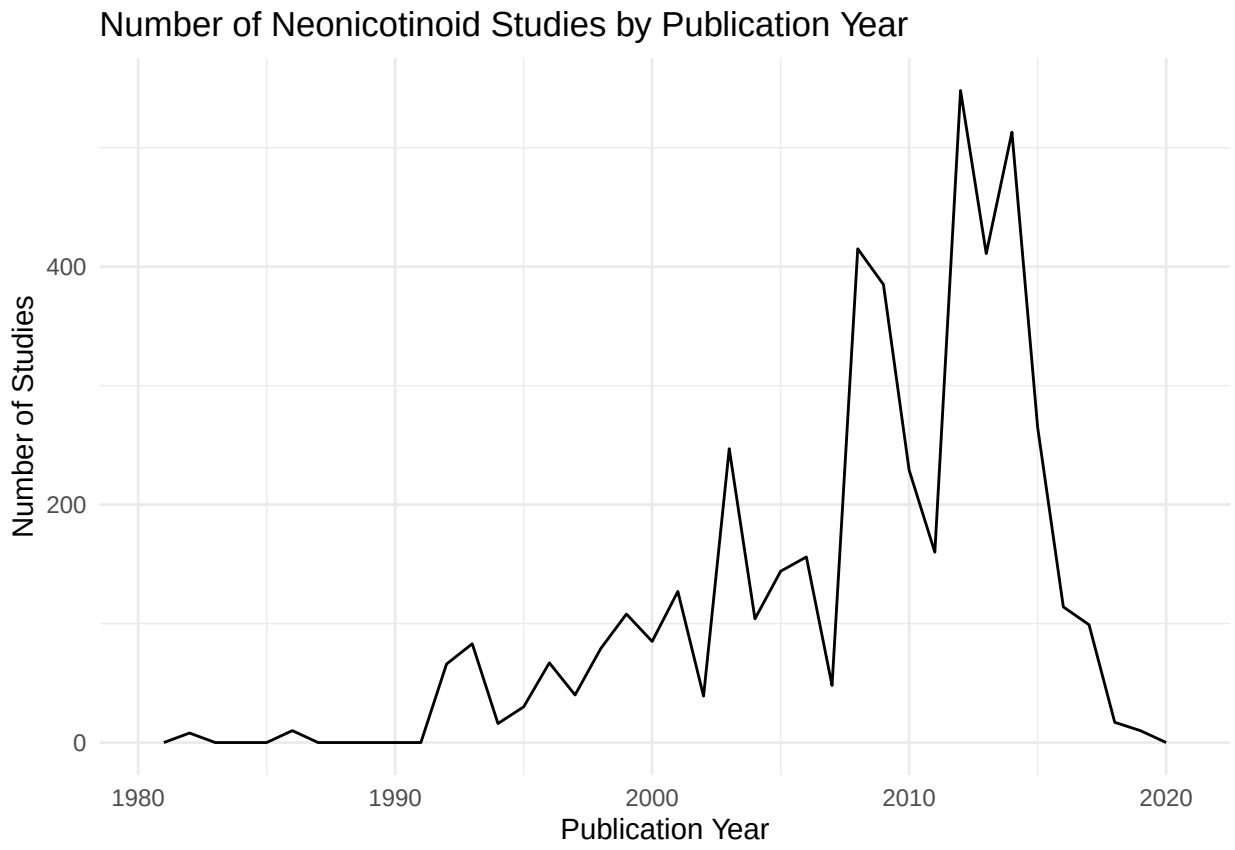
```
## [1] 27.2 19.7 47   25   13   268
## 1006 Levels: <0.0004 <0.025 <0.088 <0.5 <1.5 <10/ <2.5/ <4.00 <5.00 ... NR/
```

Answer: This is not numeric because this column represents concentrations and it is not purely numeric. From the view of dataframe, the column also includes inequality symbols, ranges, and texts. These cannot be considered as numeric data type and then R imports the entire column as a factor variable instead.

Explore your data graphically (Neonics)

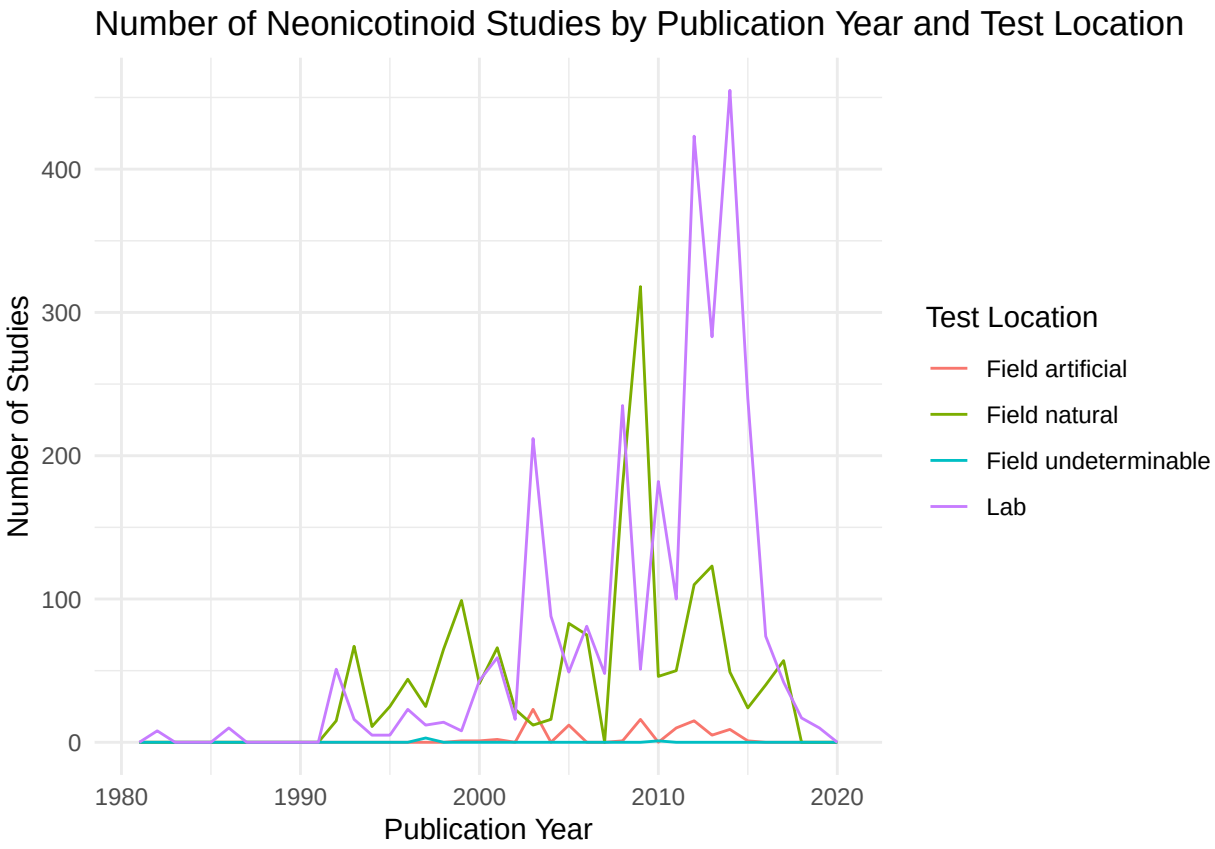
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```r
ggplot(Neonics, aes(x = Publication.Year)) +
  geom_freqpoly(binwidth = 1) +
  labs(
    x = "Publication Year",
    y = "Number of Studies",
    title = "Number of Neonicotinoid Studies by Publication Year"
  ) +
  theme_minimal()
```



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",
    color = "Test Location",
    title = "Number of Neonicotinoid Studies by Publication Year and Test Location"
  ) +
  theme_minimal()
```

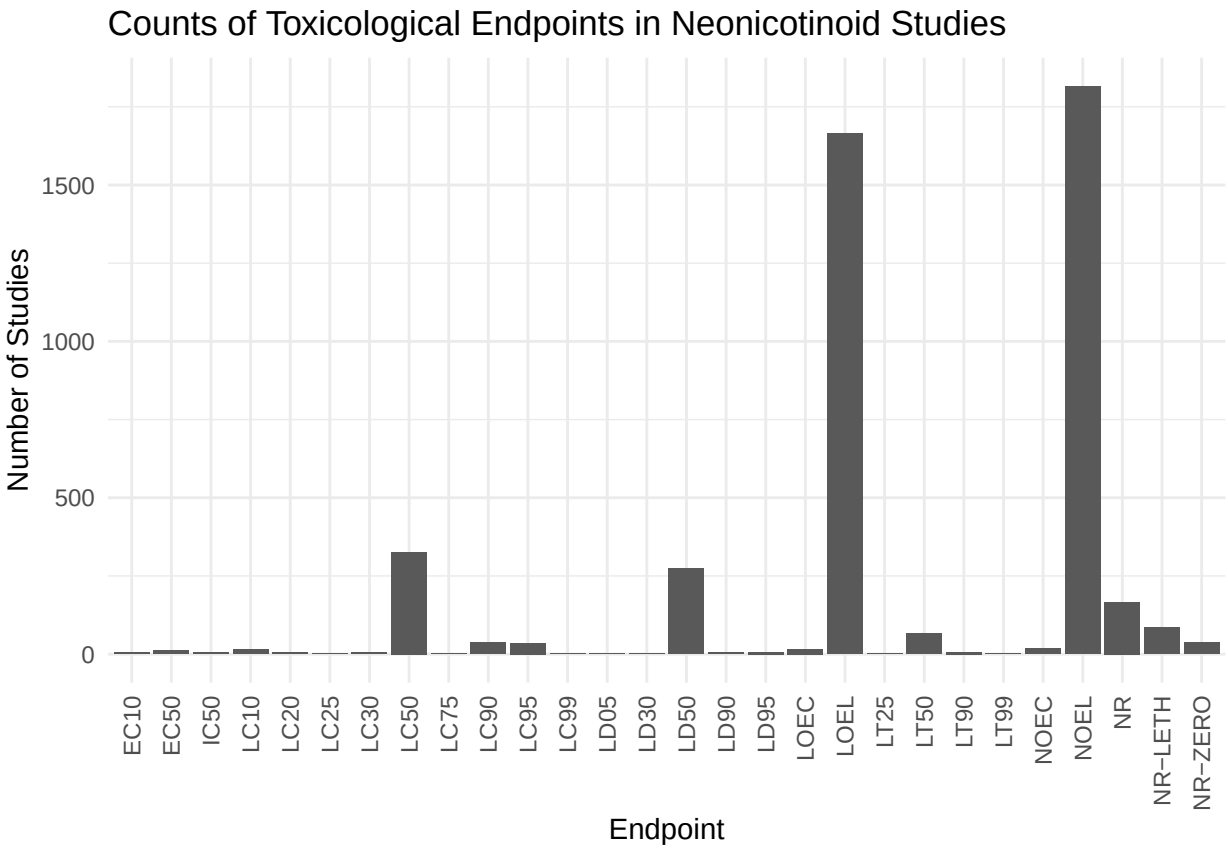


Interpret this graph. What are the most common test locations, and do they differ over time? Answer: The most common test location is lab. Even though the lab is one of the most common test locations for every decades (field under natural condition used to be dominant in several periods from 1990 to 2010), there is a trend that lab test become more frequent from 2000 to the mid of 2010.

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() +
  labs(
    x = "Endpoint",
    y = "Number of Studies",
    title = "Counts of Toxicological Endpoints in Neonicotinoid Studies"
  ) +
  theme_minimal() +
  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 end points are NOEL and LOEL. Based on the info from ECOTOX_CodeAppendix (p.721), NOEL represents "no-observable-effect-level", which means the highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test. On the other hand, LOEL represents "lowest-observable-effect-level", which means 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.

```
class(Litter$collectDate)
```

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

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

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

```
unique(
  Litter$collectDate[
    Litter$collectDate >= as.Date("2018-08-01") &
    Litter$collectDate <= as.Date("2018-08-31")
  ]
)
```

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

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

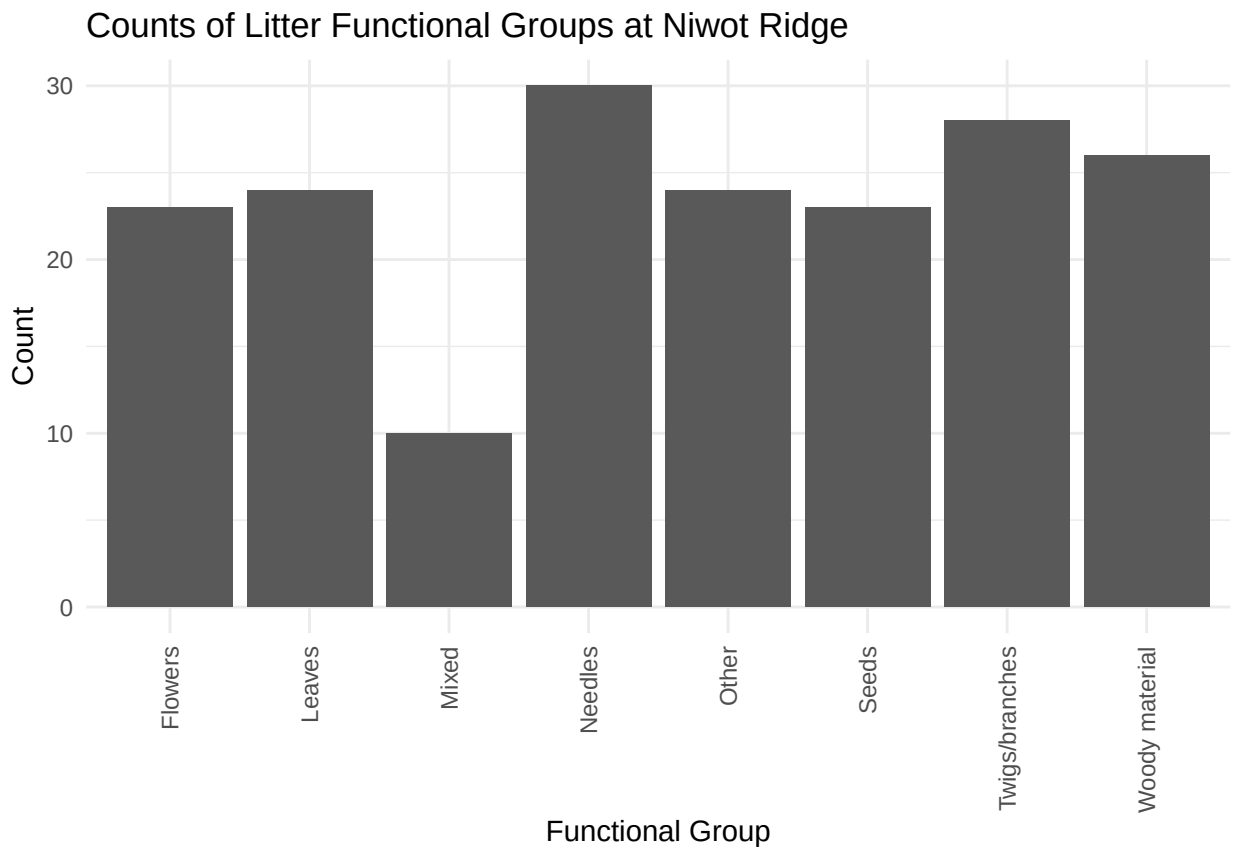
```
unique(Litter$plotID)
```

```
##  [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: The "unique" function returns all different values in a variable but not presenting the counts of values; while the "summary" function returns all different values with their frequencies to occur.

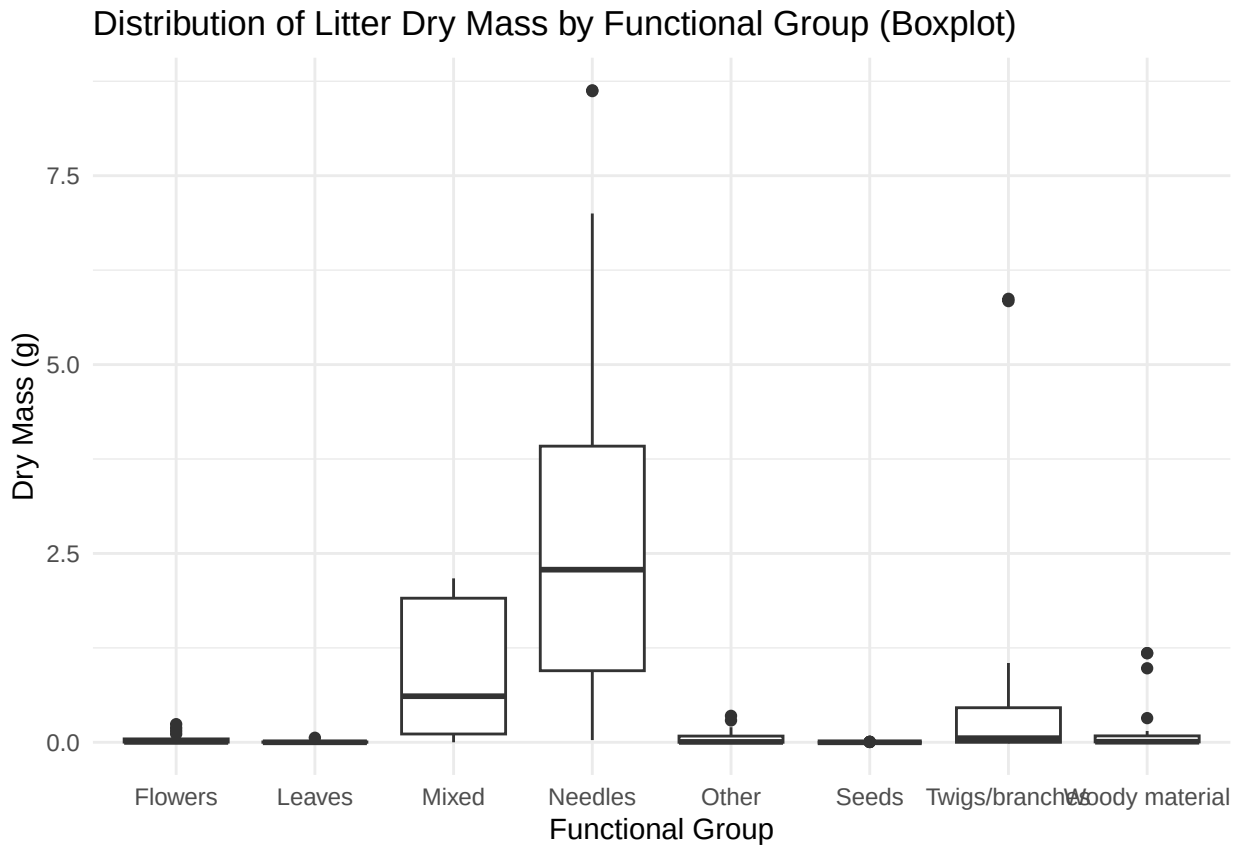
14. Create a bar graph of `functionalGroup` counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

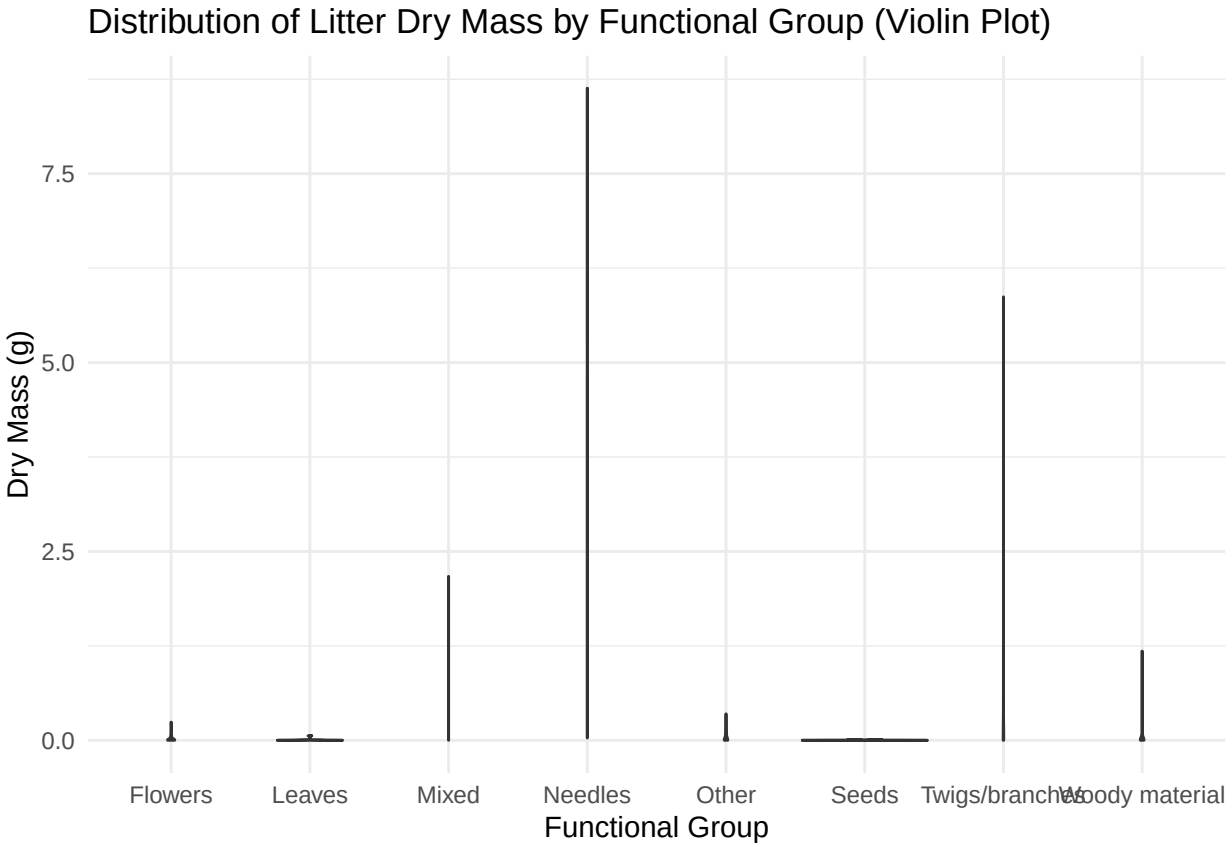


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of `dryMass` by `functionalGroup`.

```
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_boxplot() +
  labs(
    x = "Functional Group",
    y = "Dry Mass (g)",
    title = "Distribution of Litter Dry Mass by Functional Group (Boxplot)"
  ) +
  theme_minimal()
```




```r
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_violin() +
  labs(
    x = "Functional Group",
    y = "Dry Mass (g)",
    title = "Distribution of Litter Dry Mass by Functional Group (Violin Plot)"
  ) +
  theme_minimal()
```



Why is the boxplot a more effective visualization option than the violin plot in this case? Answer: The boxplot is more effective because it contains the median, the spread, and the outlier of dry mass for each functional group, while the violin plot is greatly influnced by the outliers and doesn't well present the spread of the dry mass for eah functional group.

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