Assignment 3: Data Exploration

Leah Li

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.

```r
#Load packages
library(tidyverse)
library(lubridate)
library(here)
library(ggplot2)

#Check current working directory
getwd()
```

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

```
here()
```

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

```
#Upload ECOTOX neonicotinoid dataset
Neonics.data <- read.csv(
  file = here("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"),
  stringsAsFactors = TRUE)


#Upload Niwot Ridge NEO dataset
Litter.data <- 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: We might be interested in the ecotoxicology of neonicotinoids on insects because they can negatively impact non-target species like pollinators, which could reduce biodiversity, and disrupt important ecosystem services like pollination and pest control, which are essential for agriculture and environmental health.

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: We might be interested in studying litter and woody debris in forests because they play a key role in nutrient cycling, carbon storage, and soil formation, and they provide habitat for many organisms. Understanding their dynamics helps assess forest ecosystem health and also the effects of climate change.

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.Litterfall Traps: These are elevated traps that are placed 80 cm above the ground to collect litterfall, including leaves, needles, twigs, and seeds. Fine woody debris with a diameter less than 2 cm and a length under 50 cm is collected in these traps. Ground Traps for Larger Debris: Ground traps, which are larger (3 m x 0.5 m) ones, are used to collect longer fine woody debris that falls from the forest canopy but is too large to be collected by the elevated traps. 2.Sampling Frequency: Elevated traps are sampled every two weeks during peak litterfall periods (e.g., during leaf senescence in deciduous forests) and less frequently (every 1-2 months) in evergreen forests, while ground traps are sampled once per year. 3. Spatial Sampling: Sampling occurs in designated tower plots, usually 20 m x 20 m or 40 m x 40 m, depending on the site. Trap placement can be randomized or targeted based on vegetation density.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
# Using the dim() function to check the dimensions of the dataset
dim(Neonics.data)
```

```
## [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 the summary function to generate summary for the "Effect" column
effect_summary <- summary(Neonics.data$Effect)
# Sort the effects by frequency using the sort() function
sort_effects <- sort(effect_summary)
# Show the result
print(sort_effects)
```

```
##       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 studied are population changes, mortality, and behavior. Population effects provide insights into how neonicotinoids impact insect survival and reproduction, while mortality directly measures lethal outcomes. Behavioral studies focus on how exposure disrupts activities such as their eating and mating, which can affect essential ecosystem. Together, these effects could help highlight the ecological impact of neonicotinoids on insect health and ecosystems.

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...]

```
# Generate a summary of column "Common Name" with a limit of the top 6 most frequent species
species_summary <- summary((Neonics.data$Species.Common.Name), maxsum = 6)
#Print out the result
print(species_summary)
```

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

Answer: The species such as honey bees, bumblebees, and parasitic wasps are primarily pollinators or beneficial insects in agricultural systems. They play a critical role in pollination, which is essential for food production and ecosystem health. These species are of particular interest because neonicotinoids have been linked to declines in pollinator populations like mentioned above, which can disrupt biodiversity.

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…]

```
# Check the class of the 'Conc.1..Author.' column
class(Neonics.data$Conc.1..Author.)
```

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

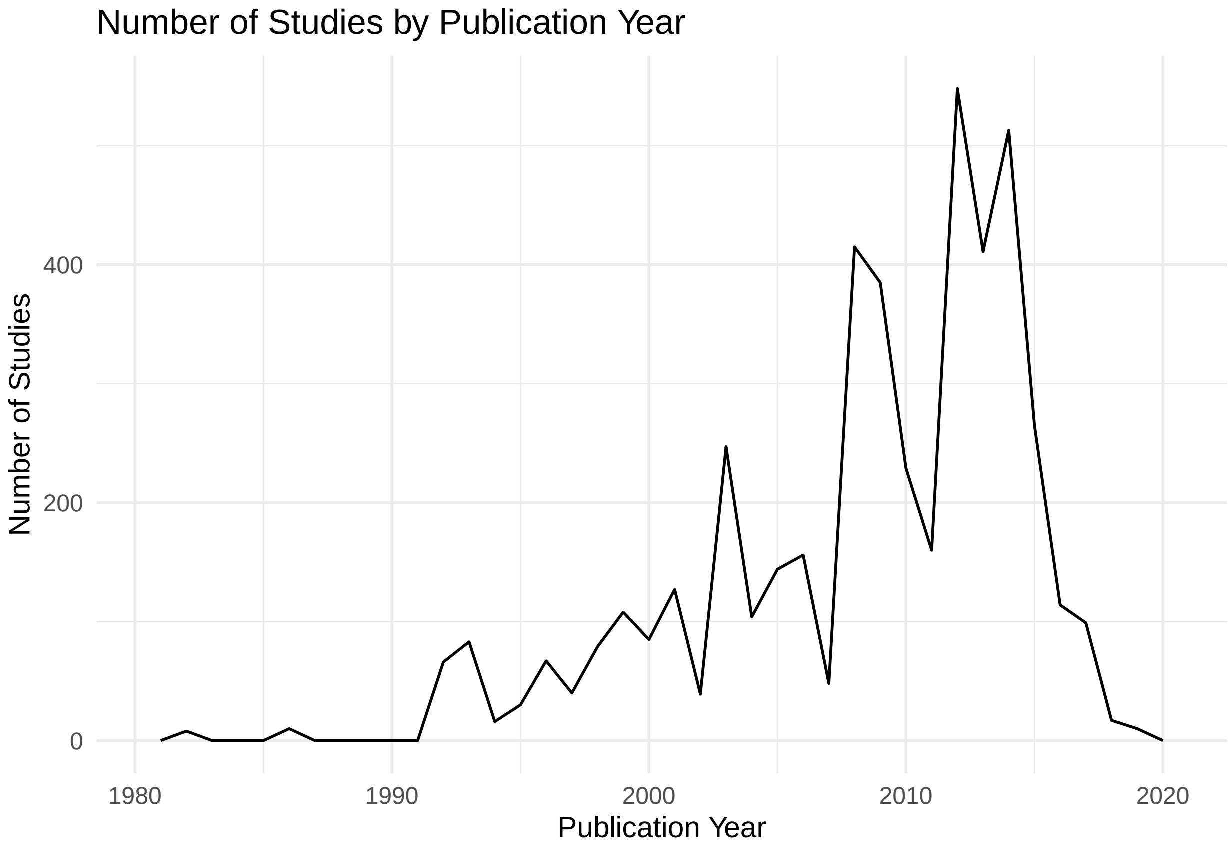
```
# View the dataframe to understand why it might not be numeric
#View(Neonics.data), here I comment it out to prevent unsuccessful knitting
```

Answer: The class of `Conc.1..Author.` column in the dataset is 'factor'. The Conc.1..Author. column is likely not numeric because it contains special characters like ">", "<", which are non-numeric symbols. These characters prevent R from recognizing the entire column as numeric.

Explore your data graphically (Neonics)

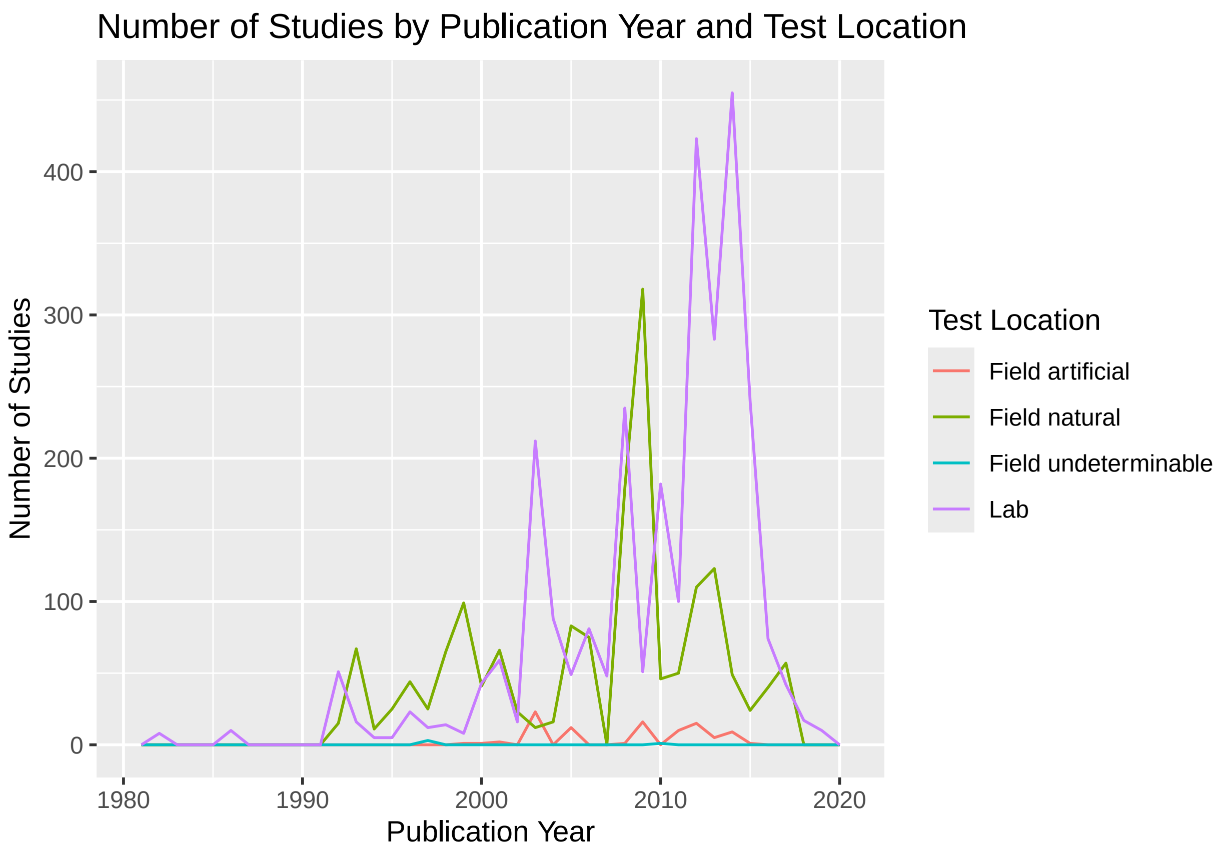
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
#Setting Publication year as x axis
ggplot(Neonics.data, aes(x = Publication.Year)) +
  geom_freqpoly(binwidth = 1) +
  labs(title = "Number of Studies by Publication Year",
       x = "Publication Year",
       y = "Number of Studies") +
  theme_minimal()
```



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

```r
#Plot Studies by Publication Year with Color by Test.Location
ggplot(Neonics.data, aes(x = Publication.Year, color = Test.Location)) +
  geom_freqpoly(binwidth = 1) +
  labs(title = "Number of Studies by Publication Year and Test Location",
       x = "Publication Year",
       y = "Number of Studies",
       color = "Test Location")  # Label for the color legend
```



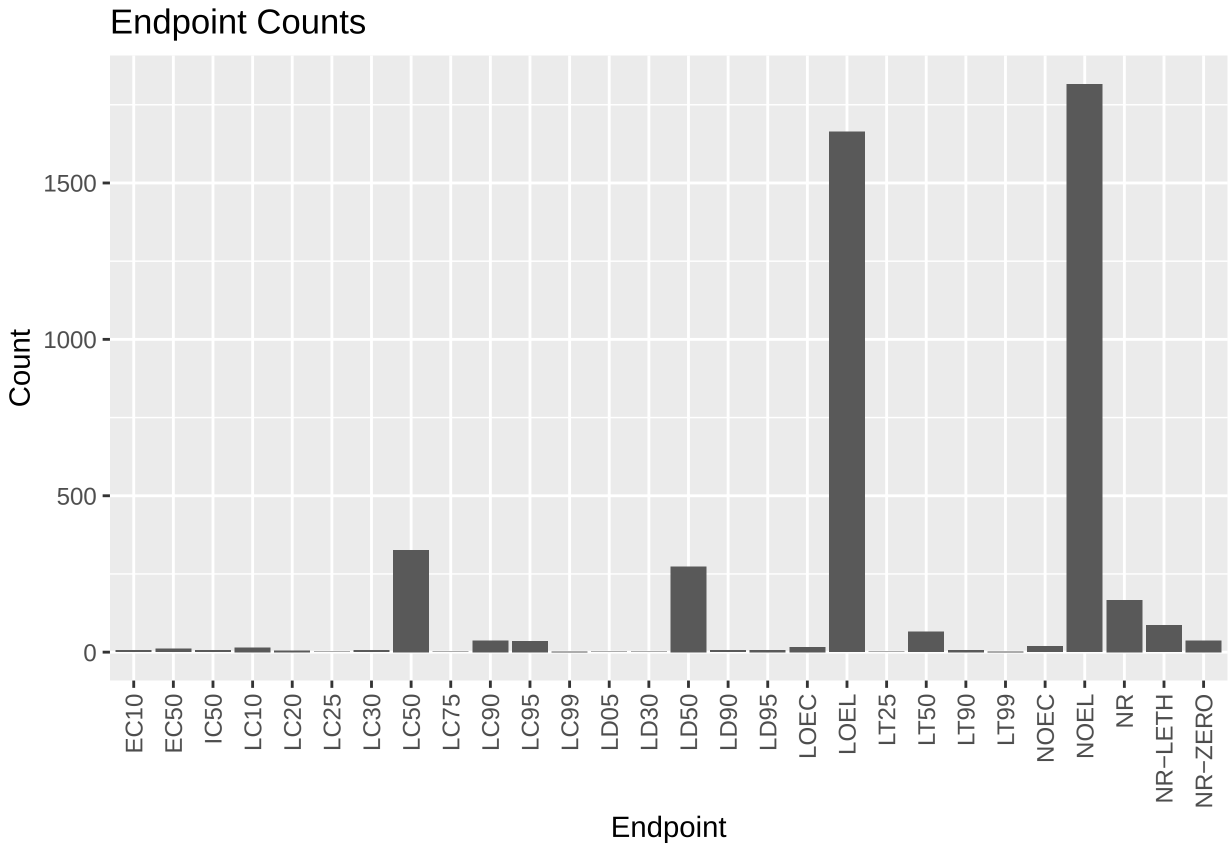
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer:The graph shows that Lab studies have been the most common test location, particularly between the late 1990s and 2010, where we observe a significant rise in the number of studies. However, after 2010, the number of lab-based studies shows a noticeable decline. Field natural (green line) is the second most common test location, with a noticeable increase in studies in the early 2000s. But it shows a decline after 2010.

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…]

```r
# --- Chunk: Plot Endpoint Counts ---
ggplot(Neonics.data, aes(x = Endpoint)) +
  geom_bar() +
  labs(title = "Endpoint Counts",
       x = "Endpoint",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer:The most two common endpoints are LOEL and NOEL. LOEL (Lowest-Observable-Effect-Level): This represents the lowest dose (concentration) at which effects were observed that are significantly different from the control group. It indicates the smallest amount of a substance that produces a measurable effect in the test organisms. NOEL (No-Observable-Effect-Level): This is the highest dose (concentration) that does not produce statistically significant effects compared to the control group. It represents the concentration below which no adverse effects are observed.

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 current class of the collectDate column
class(Litter.data$collectDate)
```

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

```
#Convert collectDate to Date
Litter.data$collectDate <- as.Date(Litter.data$collectDate, format = "%Y-%m-%d")

# Confirm the new class of collectDate
class(Litter.data$collectDate)
```

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

```
# Use the unique() function to find all dates in August 2018
august_2018 <- unique(Litter.data$collectDate[format(Litter.data$collectDate, "%Y-%m") == "2018-08"])

# Print the unique dates for August 2018
print(august_2018)
```

```
## [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`?

```
# Get the unique plots sampled at Niwot Ridge using the 'plotID' column
unique_plots <- unique(Litter.data$plotID)
# Print the number of unique plots
print(length(unique_plots))
```

```
## [1] 12
```

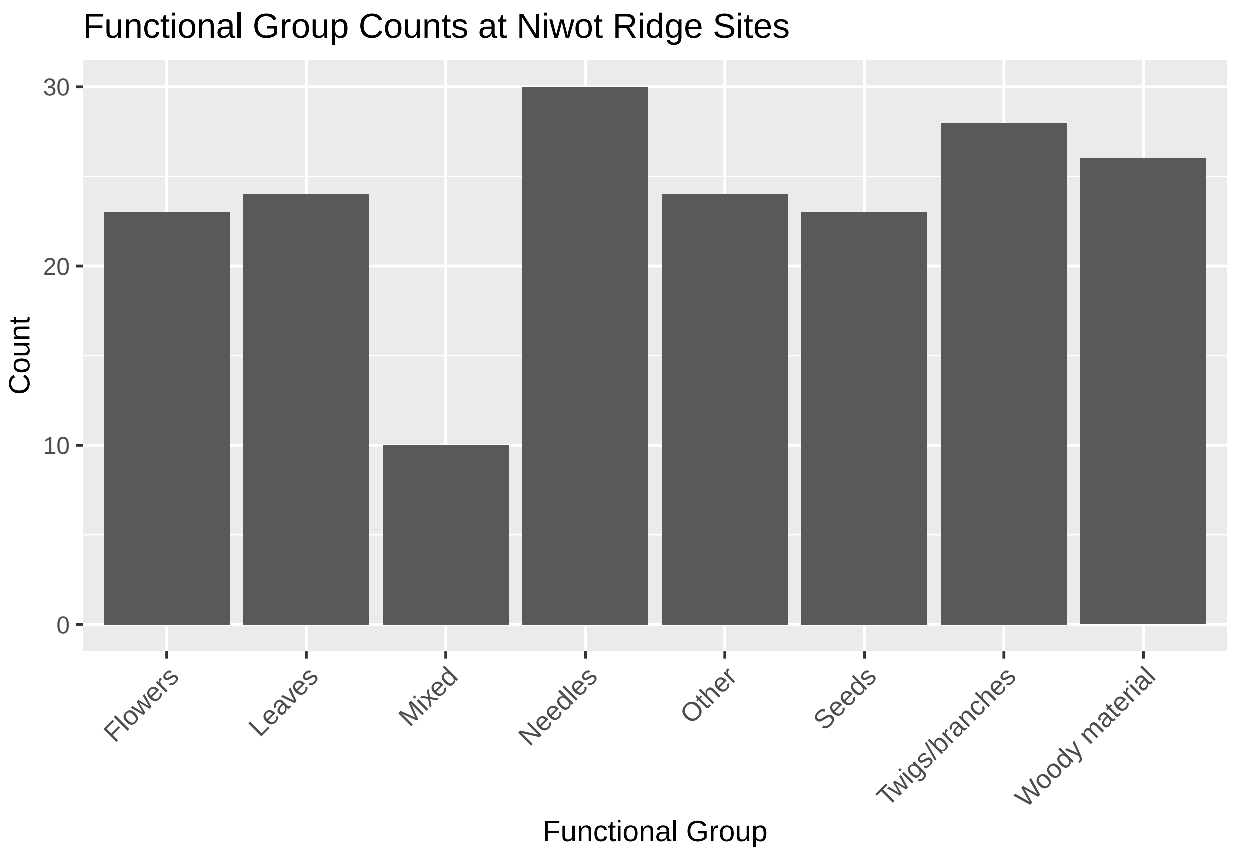
```
print(unique_plots)
```

```
##  [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
```

Answer: The unique function returns only the distinct values in a column, helping to identify how many and which specific plots were sampled at Niwot Ridge. In contrast, the summary function provides both the distinct values and their frequency of occurrence, showing how often each plot was sampled.

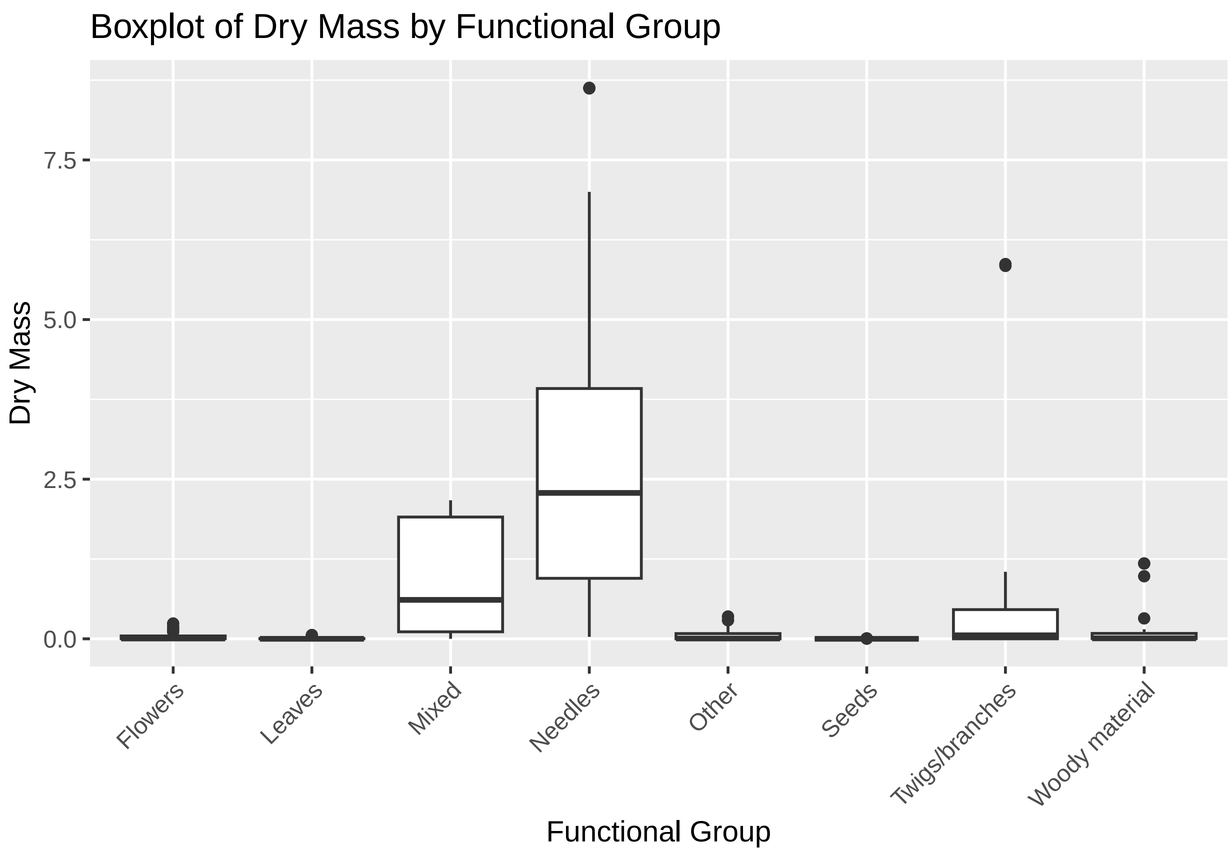
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
# Chunk: Plot Functional Group Counts
ggplot(Litter.data, aes(x = functionalGroup)) +
  geom_bar() +
  labs(title = "Functional Group Counts at Niwot Ridge Sites",
       x = "Functional Group",
       y = "Count") +
   # Adjust text angle and size
theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1, size = 10))
```

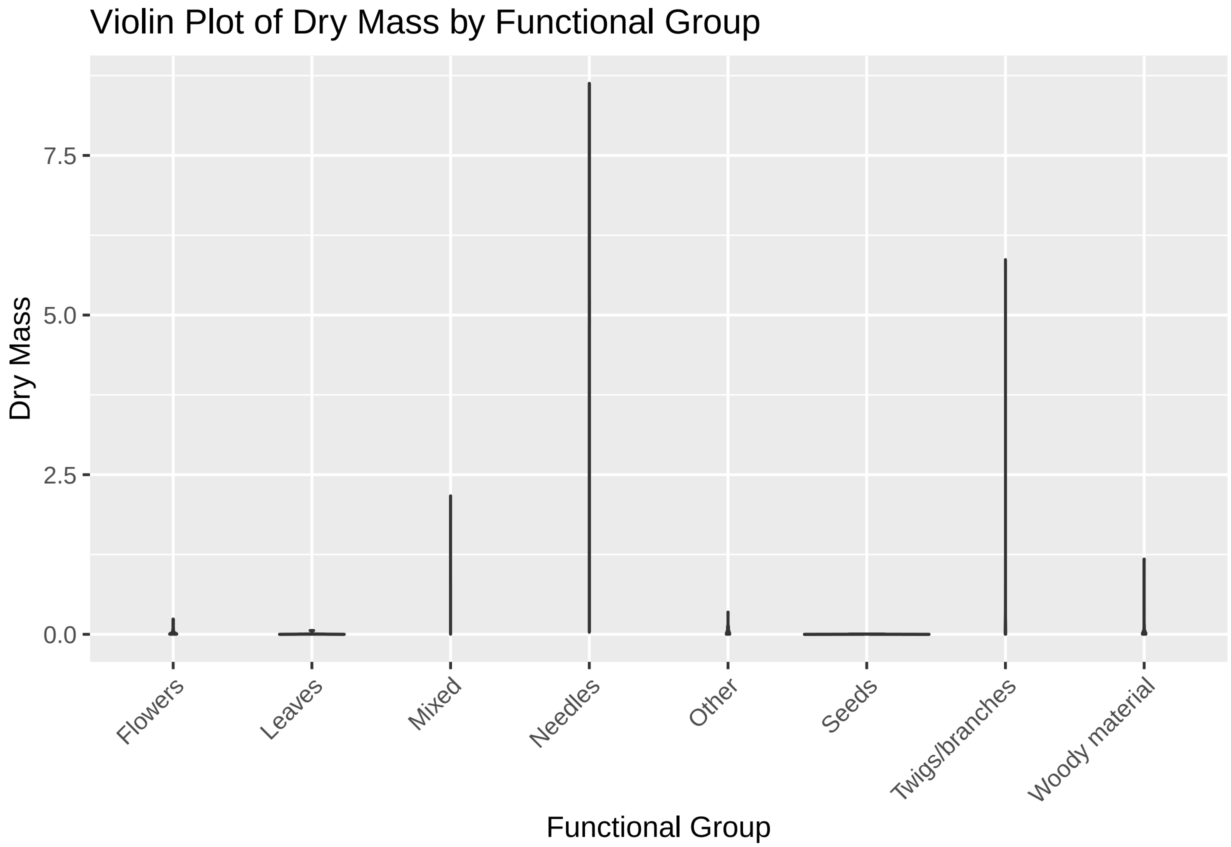


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
# Create Boxplot of dryMass by Functional Group
ggplot(Litter.data, aes(x = functionalGroup, y = dryMass)) +
  geom_boxplot() +
  labs(title = "Boxplot of Dry Mass by Functional Group",
       x = "Functional Group",
       y = "Dry Mass") +
   # Rotate x-axis labels
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```



```
# Create Violin Plot of dryMass by Functional Group
ggplot(Litter.data, aes(x = functionalGroup, y = dryMass)) +
  geom_violin() +
  labs(title = "Violin Plot of Dry Mass by Functional Group",
       x = "Functional Group",
       y = "Dry Mass") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```



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

> Answer: In this case, the boxplot is more effective than the violin plot because it clearly shows the median, quartiles, and outliers for each functionalGroup, providing a better understanding of the central tendency and variability of the dryMass data. The violin plot is less informative here due to the small sample size and less variation of the data, therefore, it has shown almost as a straight line instead of the useful violin shape.

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

> Answer: From the boxplot, the litter types that tend to have the highest biomass at these sites are Needles. The median dry mass of Needles is higher than other functional groups, with a wider range (the long vertical straight line) of variation and several outliers (dots).