Assignment 3: Data Exploration

Yufan Du

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.

```
# Load necessary libraries
library(tidyverse)
library(lubridate)
library(here)
library(ggplot2)

# Check the working directory
getwd()
```

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

```r
# Load datasets
Neonics <- read_csv(here("~/EDE_Fall2024/Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"),
                    col_types = cols(.default = "f"))

Litter <- read_csv(here("~/EDE_Fall2024/Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv"),
                   col_types = cols(.default = "f"))
```

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: Neonicotinoids are highly effective insecticides but have been associated with negative impacts on non-target insect populations, including bees and other pollinators, which are crucial for biodiversity and agriculture. (I have no idea of ecotoxicology, this all from internet.)

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: Forest litter and woody debris play a significant role in nutrient cycling, carbon storage, and habitat provision for various organisms. Woody debris is an important part of forest and stream ecosystems because it has a role in carbon budgets and nutrient cycling, is a source of energy for aquatic ecosystems, provides habitat for terrestrial and aquatic organisms, and contributes to structure and roughness, thereby influencing water flows and sediment transport (Harmon and others 1986).

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.Spatial Sampling Design: Litter and fine woody debris are sampled using elevated and ground traps at NEON sites with woody vegetation taller than 2 meters. Elevated traps are 0.5 m^2 mesh baskets positioned approximately 80 cm above the ground, collecting materials with a diameter less than 2 cm and length less than 50 cm. Ground traps are larger, rectangular areas (3 m x 0.5 m), designed to capture longer debris not effectively collected by the elevated traps 2.Temporal Sampling Design: Elevated traps are sampled frequently, especially in deciduous forest sites during leaf fall (every two weeks) and less frequently (every 1-2 months) at evergreen sites. Ground traps are sampled once per year. The temporal resolution varies depending on the vegetation type and seasonality, adjusting to account for periods of high or low litterfall production 3. Sampling Locations and Plot Design: Litterfall sampling occurs in specific plots within NEON sites, including 40 m x 40 m tower plots and 20 m x 20 m plots, depending on the vegetation structure. Traps are placed either randomly or targeted, depending on vegetation cover, to ensure representative sampling. Plot and trap locations are carefully designed to avoid interference with other ecological measurements and maintain adequate spacing between plots

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
# Check 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. Why might these effects specifically be of interest? [Tip: The `sort()` command is useful for listing the values in order of magnitude...]

```
# Summary of the Effect column
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
##     Intoxication     Accumulation     Biochemistry          Cell(s)
##               12               12               11                9
##       Physiology        Histology       Hormone(s)
##                7                5                1
```

Answer:The most common effect that is studied is population. This is probably because population is a basic information for most study.

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 of the species column with maxsum option
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
```

Answer: The six most commnoly studies speacies are honey bee, parasitic wasp buff, buff tailed bumblebee, carniolan honey bee, bumble bee and Italian honeybee. Common things: They are widely distributed, have significant roles in pollination, and are easily manageable for research. Why interest over other bees: These species are crucial for large-scale crop pollination, have unique behaviors and resilience traits, and are highly impacted by diseases and environmental changes, making them vital for studies on ecological and agricultural sustainability.

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 class of the concentration column
class(Neonics$`Conc 1 (Author)`)
```

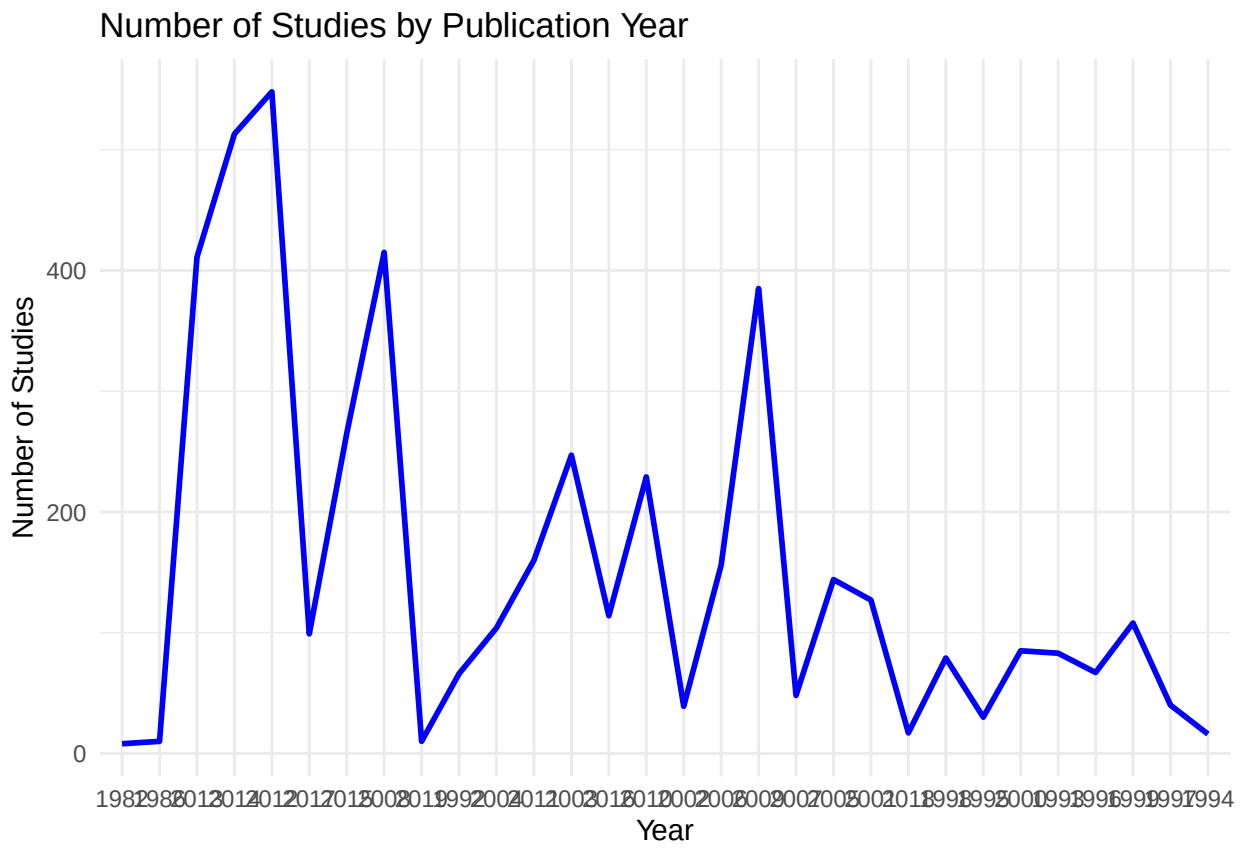
```
## [1] "factor"
```

Answer:The class is factor. It's not numeric because names are not numbers. Factors are used in R to represent categorical data, which occurs when the column includes values that cannot be directly converted to numbers.

Explore your data graphically (Neonics)

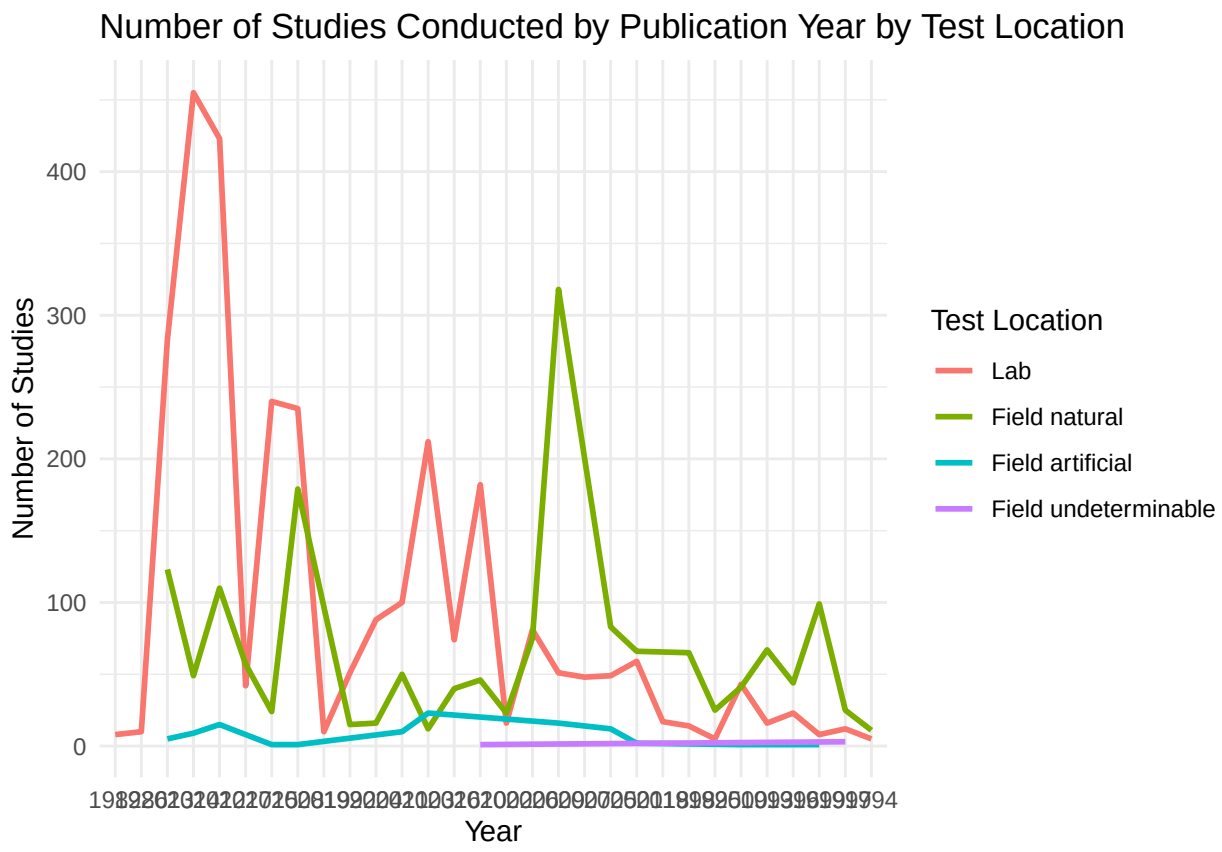
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
# Line graph of publication year
ggplot(Neonics, aes(x = `Publication Year`, group = 1)) +
  geom_freqpoly(stat = "count", color = "blue", linewidth = 1) +
  labs(title = "Number of Studies by Publication Year",
       x = "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.

```
# More color added
ggplot(Neonics, aes(x = `Publication Year`, color = `Test Location`, group = `Test Location`)) +
  geom_freqpoly(stat = "count", linewidth = 1) +
  labs(title = "Number of Studies Conducted by Publication Year by Test Location",
       x = "Year",
       y = "Number of Studies") +
  theme_minimal()
```



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

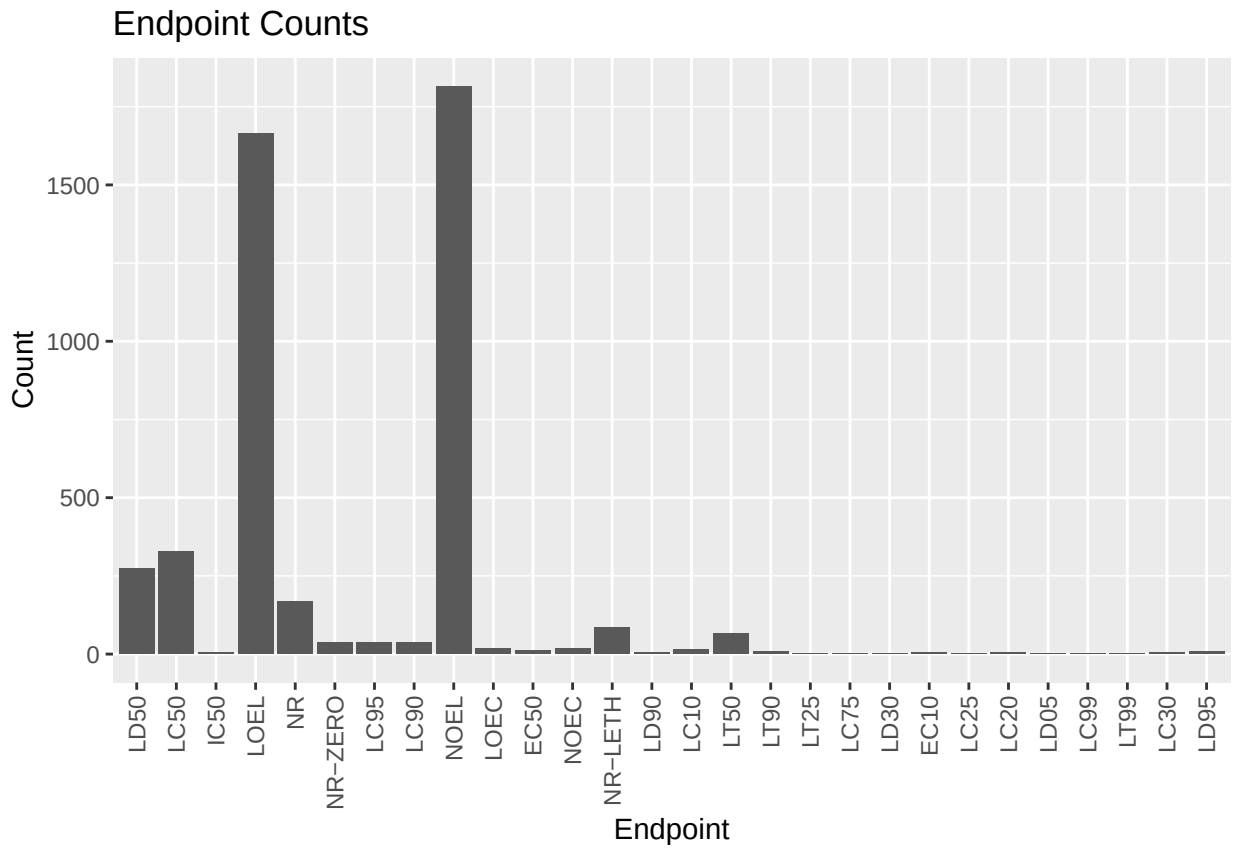
> Answer: The most common location was lab. And then after 2002, the most common location changed to firled natural. They differ over time.

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...]

```
# 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)) +
  labs(title = "Endpoint Counts", x = "Endpoint", y = "Count")
```



Answer: NOEL and LOEL. LOEL means Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC). NOEL means 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).

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 and convert collectDate to date class
class(Litter$collectDate)
```

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

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

```
## [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`?

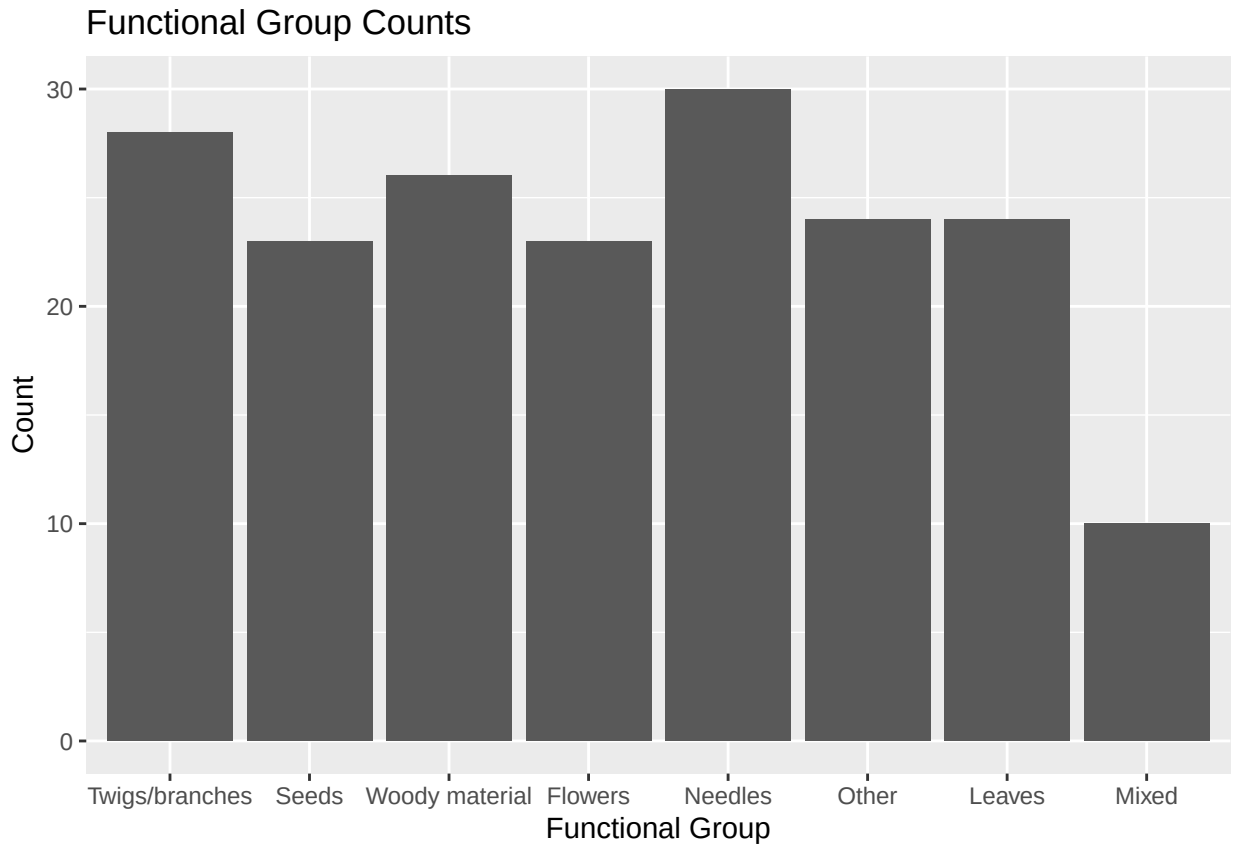
```
# Unique plots
unique(Litter$plotID)
```

```
##  [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_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 ... NIWO_057
```

Answer: Unique lists every polt. And summary adds everything together and provides an overview of counts.

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.

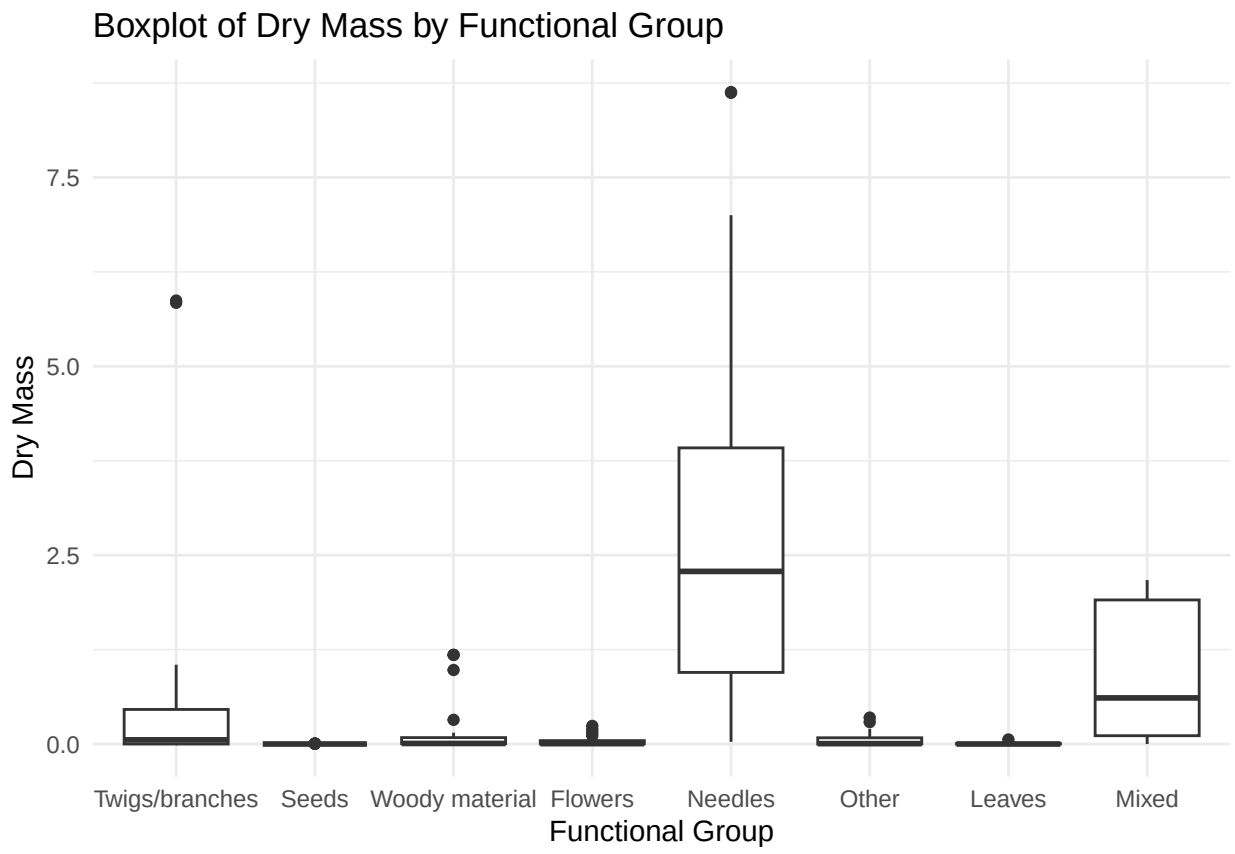
```
# Bar graph of functional groups
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar() +
  labs(title = "Functional Group Counts", x = "Functional Group", y = "Count")
```



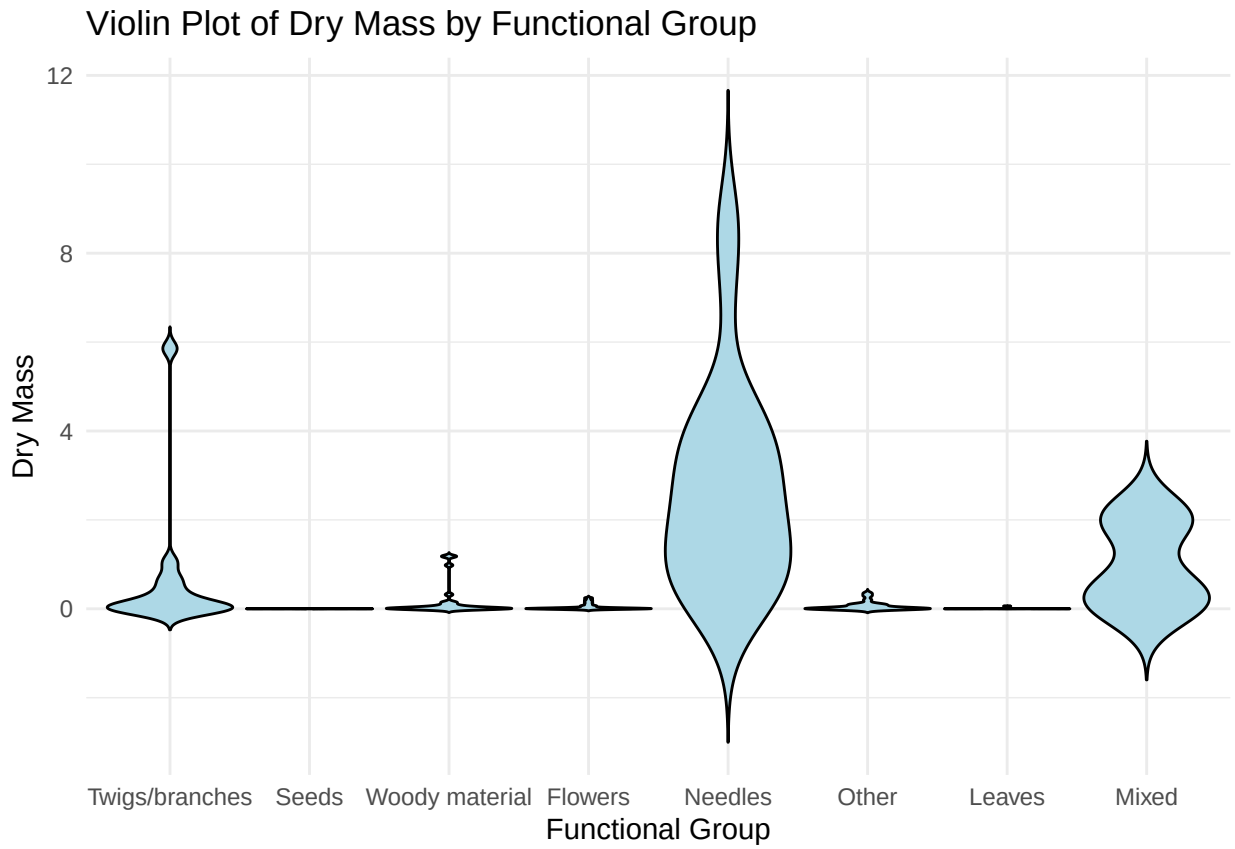
15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```r
# Covert data to numeric
Litter$dryMass <- as.numeric(as.character(Litter$dryMass))

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




```r
# Boxplot graph of functional groups
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_violin(trim = FALSE, fill = "lightblue", color = "black", scale = "width") + # Ensures the tails
  labs(title = "Violin Plot of Dry Mass by Functional Group", x = "Functional Group", y = "Dry Mass") +
  theme_minimal()
```



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

Answer: This is because Violin plot only has a shape. But boxplot shows median, quartiles, and potential outliers.

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

Answer: Needles