Assignment 3: Data Exploration

Lu Liu

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 necessary packages with library
library(tidyverse)
library(lubridate)
library(here)

#check my directory
getwd()
```

```
## [1] "/home/guest/EDA_Spring2025_ForkCeleste"
```

```r
#upload two datasets, name them, and read string as factors
neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",stringsAsFactors = TRUE)
litter <- read.csv("./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 are interested in the ecotoxicology of neonicotinoids on insects because these chemicals are widely used in agriculture and can have significant impacts on non-target insect populations, including pollinators like bees, which are crucial for ecosystem health and food production. Understanding their toxicological effects helps assess risks to biodiversity, ecosystem services, and agricultural sustainability. Additionally, studying neonicotinoids can inform regulatory decisions and the development of safer pest management practices.

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 are interested in studying litter and woody debris in forests because they play a critical role in nutrient cycling, carbon storage, and soil formation, which are essential for maintaining forest ecosystem health. This organic material also provides habitat and resources for a variety of organisms, including decomposers, insects, and microorganisms, contributing to biodiversity. Additionally, understanding litter and woody debris dynamics helps assess the impacts of climate change, forest management practices, and disturbances like wildfires on ecosystem processes.

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 frequency varies by site, ranging from monthly to annually, with deciduous forests sampled more often during leaf fall. Collected litter is sorted into categories such as leaves, twigs, seeds, and other functional groups. 2.Litter is gathered using 0.5 m^2 PVC baskets placed 80 cm above the ground, while fine woody debris is collected from 3 m x 0.5 m rectangular traps on the ground. 3.In forested areas, traps are placed randomly within a grid of potential locations, whereas in non-forested areas, traps are strategically positioned under woody vegetation.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
#use dim to know the dimensions
dim(neonics)
```

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

```
#4623 is the number of rows in the dataset
#30 is the number of columns in the dataset
```

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...]

```
#name the function and print the result
effect_counts <- summary(neonics$Effect)
print(effect_counts)
```

```
##     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 the result to get a clearer picture of which effects are
#more frquently studied and their relative importance
sort(effect_counts)
```

```
##       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: This kind of research is crucial because it informs better regulations and paves the way for safer alternatives to neonicotinoids. Neonicotinoids have far-reaching effects on insects, many of which directly impact ecosystems. For instance, increased mortality and population decline are clear signs of their harmful influence, particularly on pollinators and other beneficial insects that play vital roles in nature and agriculture. But it's not just about immediate death—sublethal effects can be just as damaging. Things like reduced feeding, reproductive issues, and developmental delays might not kill insects outright, but they can weaken populations over time, leading to long-term declines. On a deeper level, less obvious effects, such as changes in biochemistry, genetics, and immune responses, help scientists uncover how these chemicals work inside organisms.

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...]

```r
#summarize the species
#sort the result and name it
#sort the top 6 out
summary_common <- summary(neonics$Species.Common.Name)
sort_common <- sort(summary_common, decreasing=TRUE)
top_species <- head(sort_common, n=6)
print(top_species)
```

```
##              (Other)             Honey Bee        Parasitic Wasp
##                  670                   667                   285
## Buff Tailed Bumblebee   Carniolan Honey Bee            Bumble Bee
##                  183                   152                   140
```

Answer: The six most commonly studied species in the datasets are honey bee, parasitic wasp, buff tailed bumblebee, carniolan honey bee, bumble bee, italian honeybee; Most of these species are critical pollinators; these species might be of interest over other insects because many crops, such as fruits and vegeatbles rely heavily on bee pollination. The loss of these species could severely disrupt food supply chains.

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…]

```r
#check the class of the column
class(neonics$Conc.1..Author.)
```

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

```r
#view the first few rows of the column to inspect its contents
head(neonics$Conc.1..Author.)
```

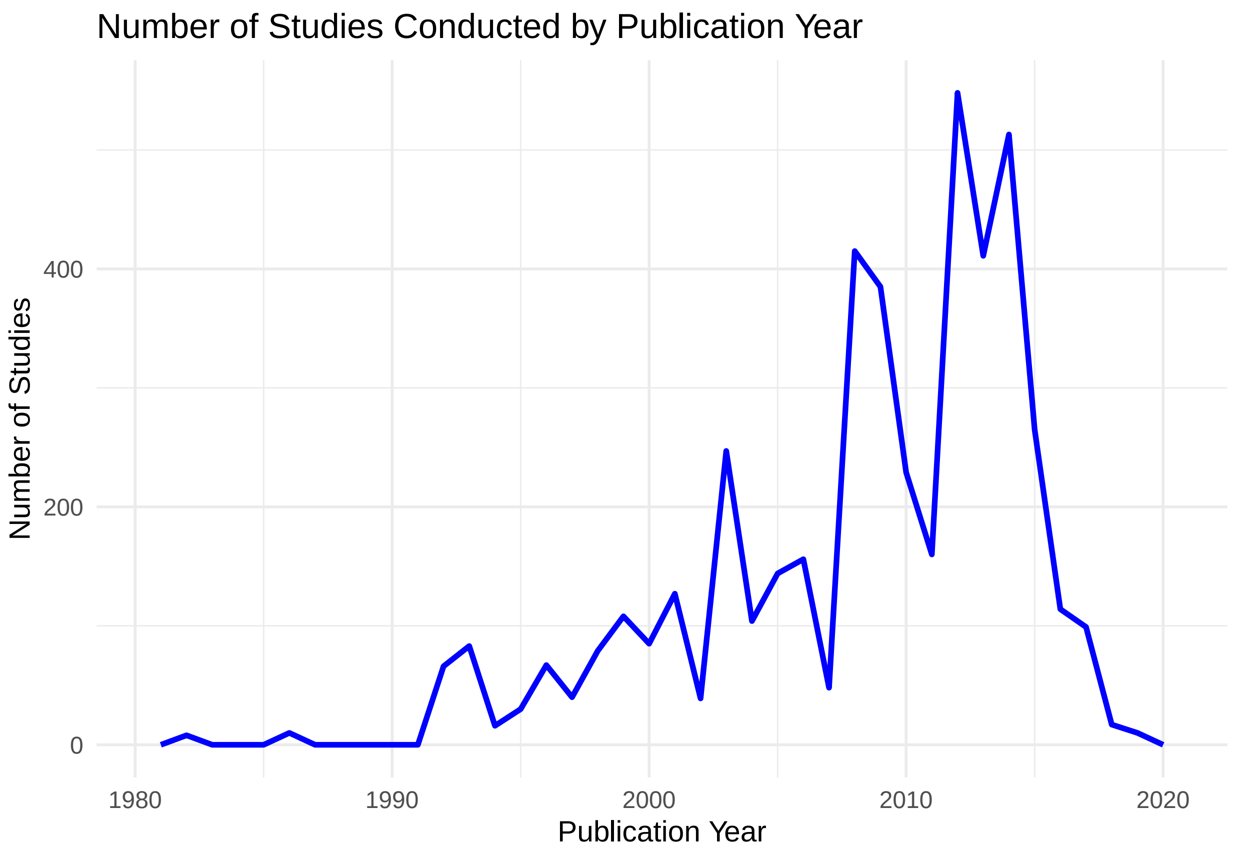
```
## [1] 27.2 19.7 47   25   13   268
## 1006 Levels: <0.0004 <0.025 <0.088 <0.5 <1.5 <10/ <2.5/ <4.00 <5.00 ... NR/
```

Answer: The class of the column is numeric, because symbols like <, >, or text like ND prevents R from interpreting the column as purely numeric.

Explore your data graphically (Neonics)

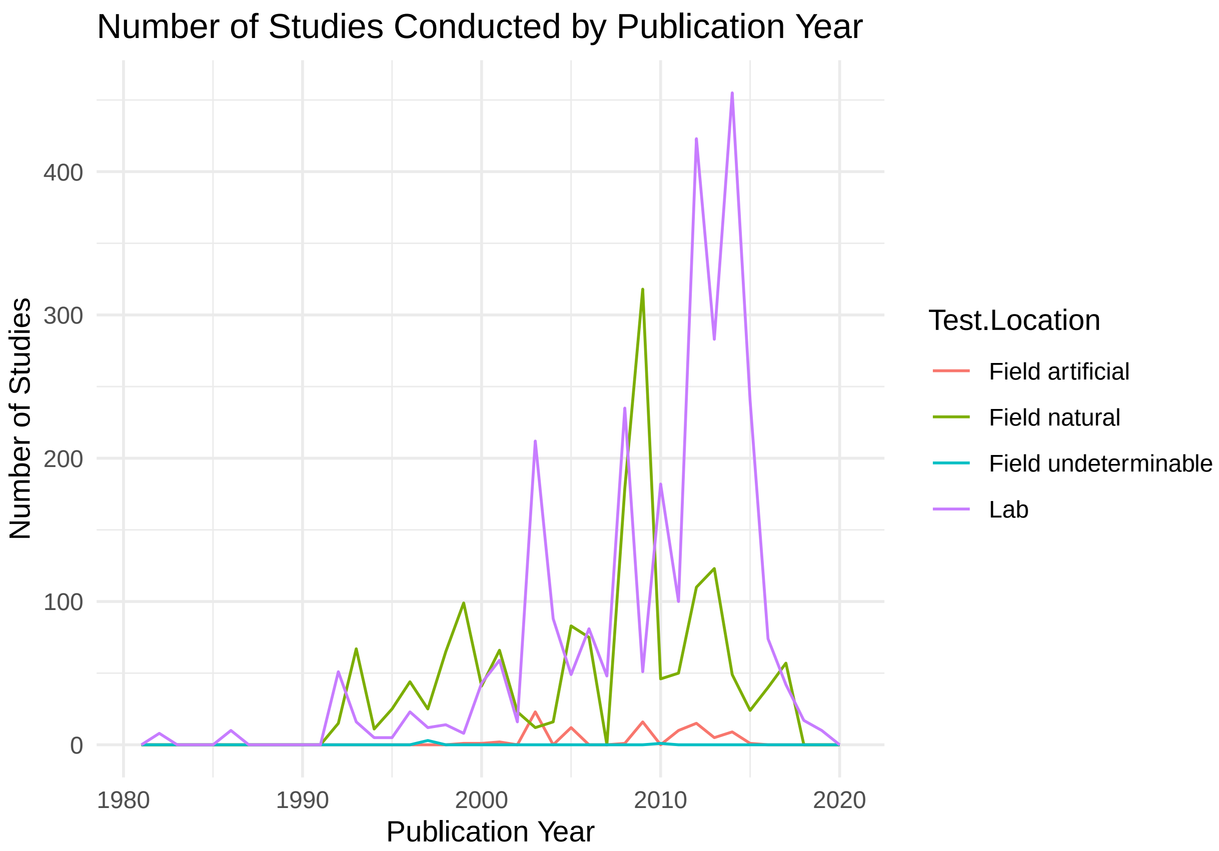
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```r
#load ggplot, use the function to generate the result
library(ggplot2)
ggplot(neonics, aes(x = Publication.Year)) +
  geom_freqpoly(binwidth = 1, color = "blue", linewidth = 1) +
  labs(
    title = "Number of Studies Conducted 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.

```
#add a color aesthetic
ggplot(neonics, aes(x = Publication.Year, color=Test.Location)) +
  geom_freqpoly(binwidth = 1) +
  labs(
    title = "Number of Studies Conducted by Publication Year",
    x = "Publication 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 frequent testing location is the lab, with the number of lab-based studies rising sharply over the years. Field Natural is the second most common location, peaking just before 2010, after which the number of studies begins to decrease.

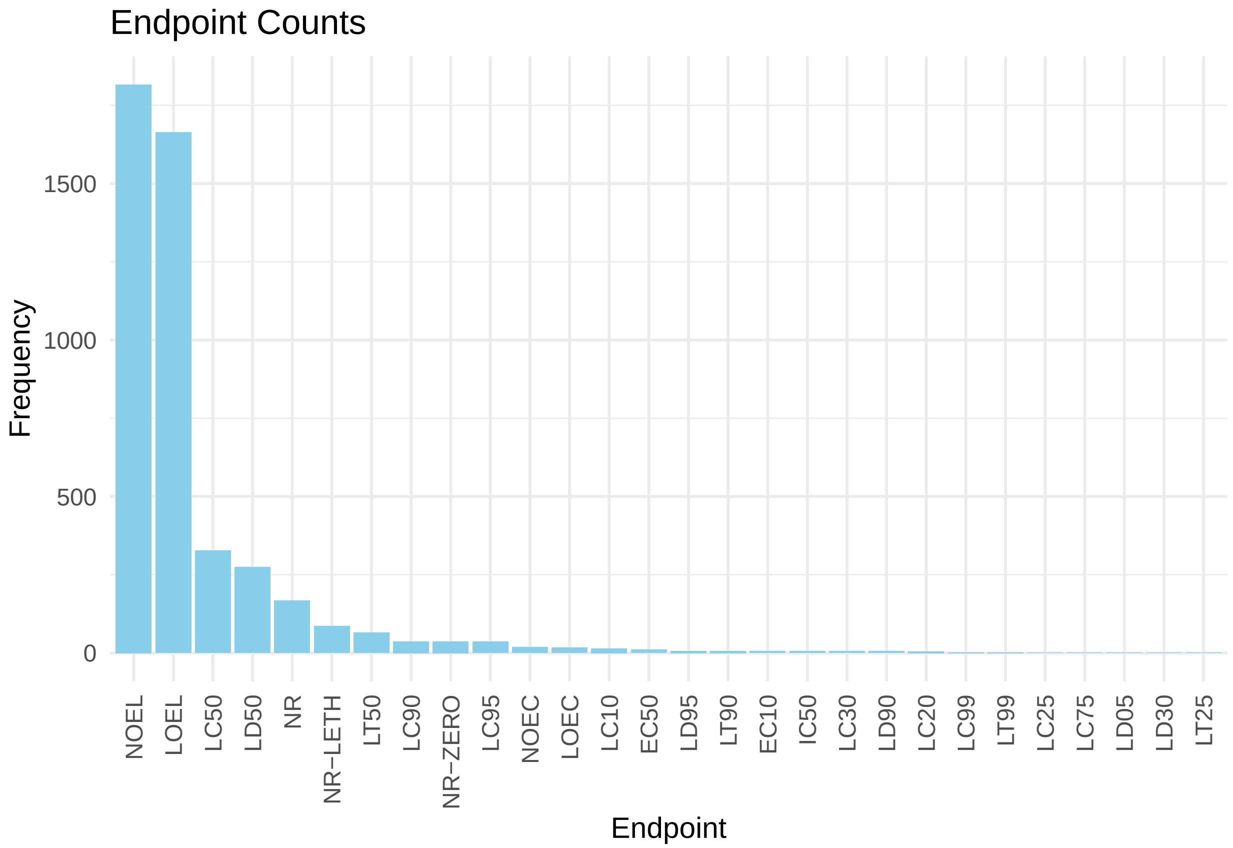
11. Create a bar graph of Endpoint counts. What are the two most common end points, and how are they defined? Consult the ECOTOX_CodeAppendix for more information.

[**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
#load the package
library(dplyr)

#calculate the count of each unique endpoint
count_endpoint <- table(neonics$Endpoint)
print(count_endpoint)
```

```
##
##    EC10    EC50    IC50    LC10    LC20    LC25    LC30    LC50    LC75    LC90
##       6      11       6      15       5       1       6     327       1      37
##    LC95    LC99    LD05    LD30    LD50    LD90    LD95    LOEC    LOEL    LT25
##      36       2       1       1     274       6       7      17    1664       1
##    LT50    LT90    LT99    NOEC    NOEL      NR NR-LETH NR-ZERO
##      65       7       2      19    1816     167      86      37
```

```r
#convert the result to data.frame
count_endpoint_df <- data.frame(
Endpoint = names(count_endpoint), Frequency = as.numeric(count_endpoint))

#sort by Frequency in descending order
sorted_endpoint_df <- count_endpoint_df %>%
  arrange(desc(Frequency))

#use gem_bar to plot the graph
ggplot(sorted_endpoint_df, aes(x = reorder(Endpoint, -Frequency), y = Frequency)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  theme_minimal() +
  labs(title = "Endpoint Counts",
       x = "Endpoint",
       y = "Frequency") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



Answer: the two most common endpoints are NOEL and LOEL. NOEL (No Observed Effect Level) acts as a safety benchmark, representing the highest dose or concentration of a substance at which no noticeable harm or significant changes are observed in the organisms under study. On the other hand, LOEL (Lowest Observed Effect Level) is the smallest dose or concentration at which detectable adverse effects or significant changes in the organisms are observed during a specific exposure period. LOEL marks the threshold where harmful effects first become apparent, indicating that the substance is causing damage or notable alterations in the organisms.

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.

```
#determine the class of collectDate
class(litter$collectDate)
```

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

```
#convert collectDate from factor to date
litter$collectDate <- as.Date(litter$collectDate)
class(litter$collectDate)
```

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

```
#now shows as date

#use the unique function to identify dates
aug_dates <- unique(litter$collectDate)
aug_2018_dates <- aug_dates[format(aug_dates, "%Y-%m") == "2018-08"]
aug_2018_dates
```

```
## [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`?

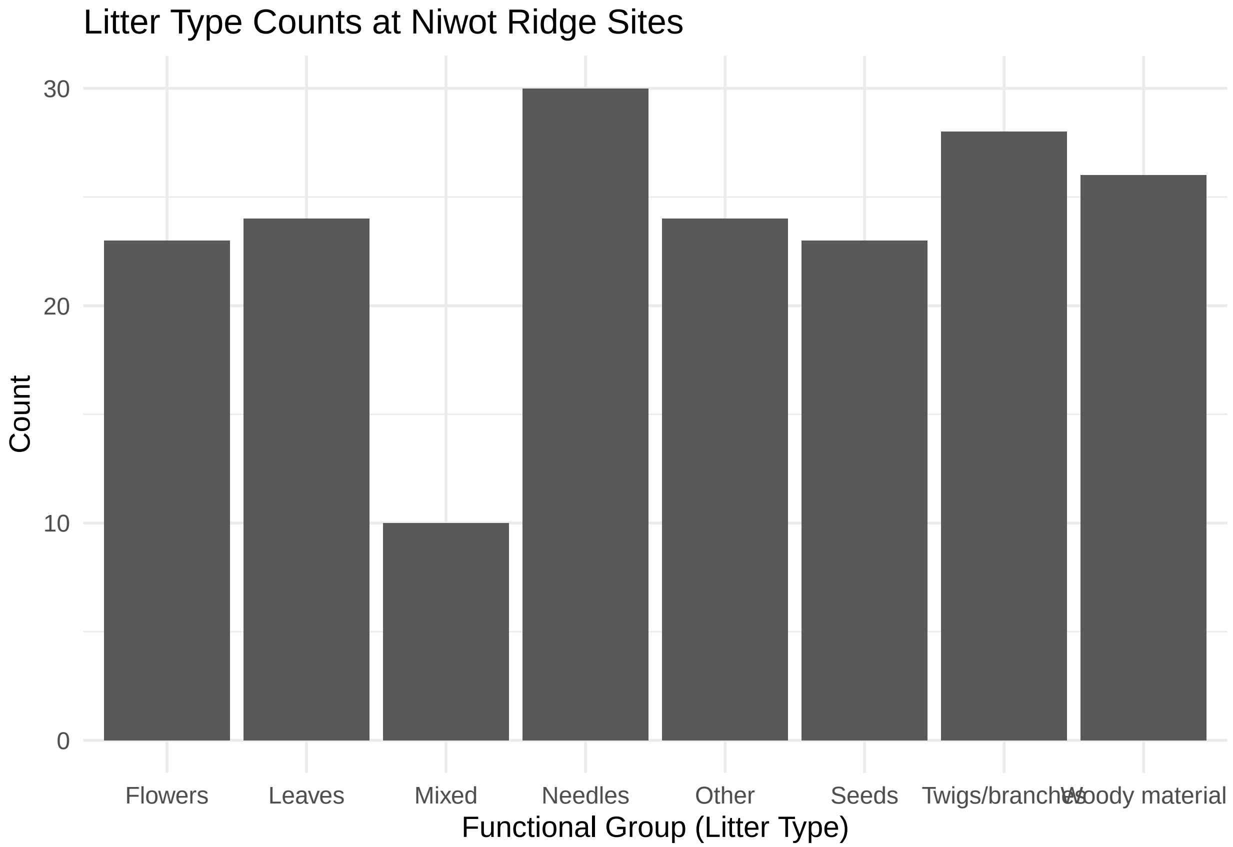
```
#find how many different plots were sampled at Niwot Ridge
unique_plots <- unique(litter$plotID)
#calculate the number of unique plot IDs
num_unique_plots <- length(unique_plots)
#display the number of unique plots sampled
num_unique_plots
```

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

> Answer: unique tells you which unique values exist in the data; summary gives a broader statistical summary, including how many times each unique value occurs (if categorical), or key statistics for numeric data.

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 the graph with ggplot
ggplot(litter, aes(x = functionalGroup)) +
  geom_bar() +
  labs(title = "Litter Type Counts at Niwot Ridge Sites",
       x = "Functional Group (Litter Type)",
       y = "Count") +
  theme_minimal()
```

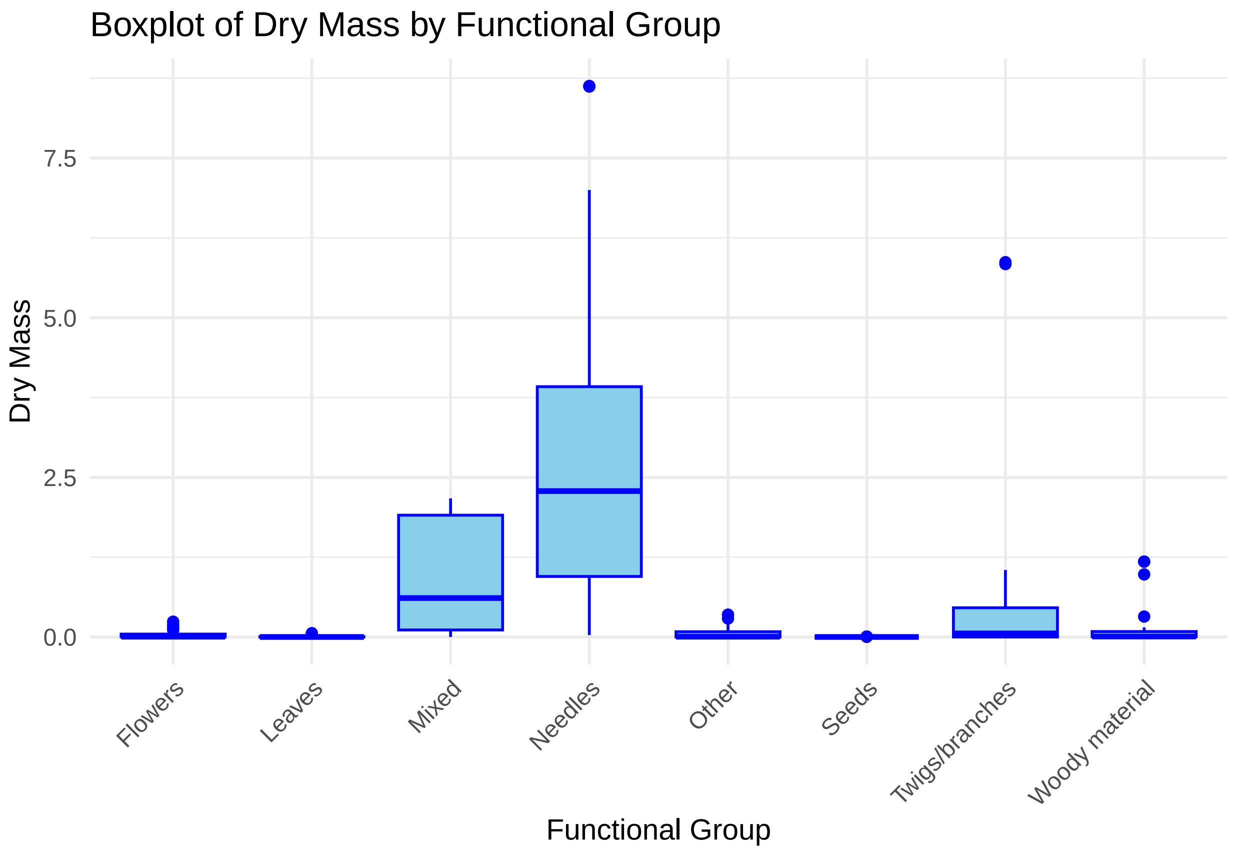


```
  theme(axis.text.x = element_text(angle = 80, vjust = 0.8, hjust = 2))
```

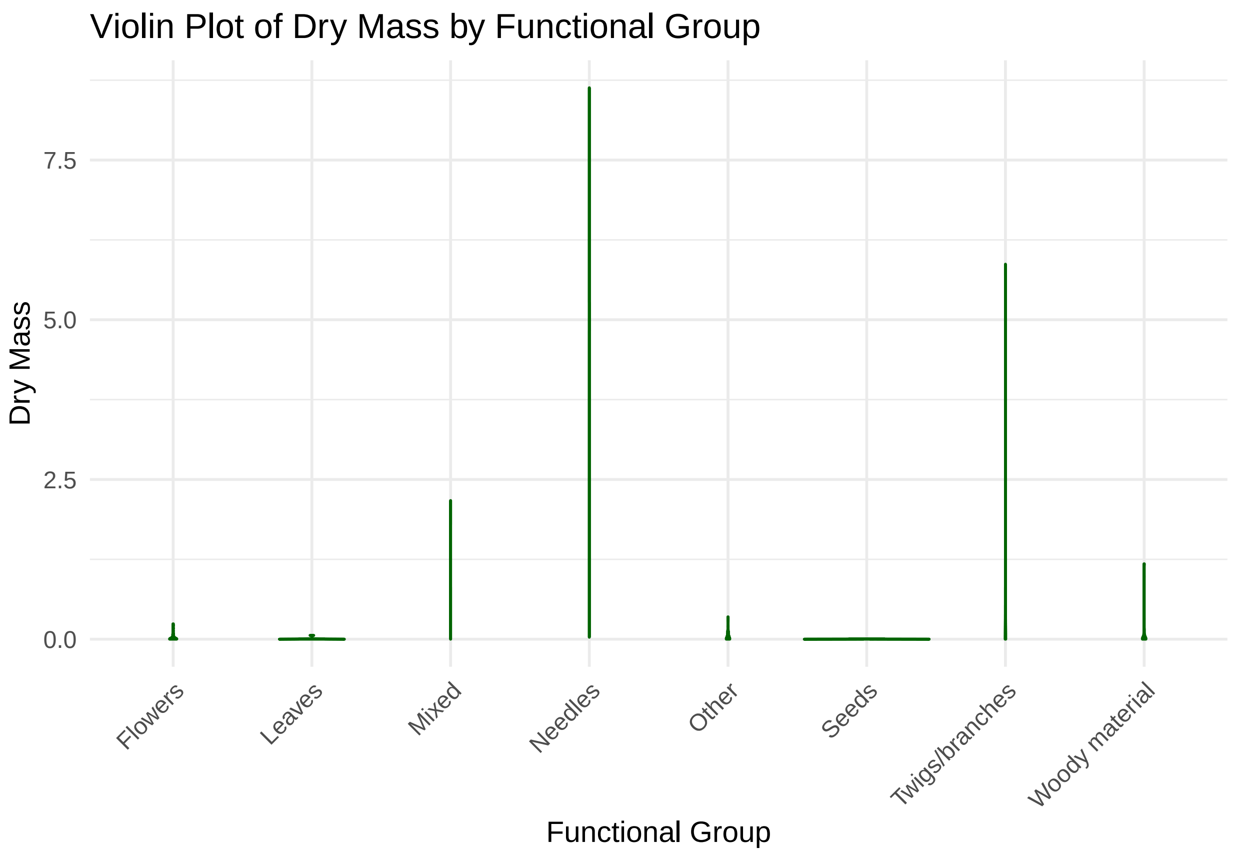
```
## List of 1
##  $ axis.text.x:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 2
##   ..$ vjust        : num 0.8
##   ..$ angle        : num 80
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

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

```r
#boxplot of dryMass by FunctionalGroup
ggplot(litter, aes(x = functionalGroup, y = dryMass)) +
  geom_boxplot(fill = "skyblue", color = "blue") +
  labs(
    x = "Functional Group",
    y = "Dry Mass",
    title = "Boxplot of Dry Mass by Functional Group"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```r
#Violin plot of dryMass by Functional group
ggplot(litter, aes(x = functionalGroup, y = dryMass)) +
  geom_violin(fill = "lightgreen", color = "darkgreen") +
  labs(
    x = "Functional Group",
    y = "Dry Mass",
    title = "Violin Plot of Dry Mass by Functional Group"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



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

> Answer: In this scenario, the boxplot is more effective than the violin plot for visualizing the distribution of dry mass values across functional groups. The boxplot clearly displays the central tendency (with the median indicated by the horizontal line), the spread (represented by the height of the box, which shows the interquartile range), and any outliers (visible as individual data points outside the box). On the other hand, the violin plot primarily highlights the median, which is already provided by the boxplot, making it less informative in this context.

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

> Answer: The group with the highest median dry mass, which is Needles.