Assignment 3: Data Exploration

Lauren Xu

Spring 2025

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 packages
library(tidyverse)
library(lubridate)
library(here)

#check the working directory
getwd()
```

```
## [1] "/home/guest/EDE_Spring2025"
```

```
here()
```

```
## [1] "/home/guest/EDE_Spring2025"
```

```
#upload the datasets "Neonics"
Neonics <- read.csv(
  file=here("/home/guest/EDE_Spring2025/Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"),
  stringsAsFactors = TRUE)

#upload the datasets "Litter"
Litter <- read.csv(
  file=here("/home/guest/EDE_Spring2025/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: Studying neonicotinoid ecotoxicology is essential to balance agricultural productivity with environmental sustainability. The ECOTOX dataset provides evidence to mitigate unintended harms, safeguard ecosystems, and ensure long-term food security.

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: The data helps track climate change effects on alpine ecosystems and informs sustainable forest management.

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.spacial sampling 2.temporal sampling 3.trap types and materials collected

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
# read the data "Neonics"
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...]

```
# Generate a summary of the "Effect" column
summary(Neonics$Effect)
```

```
##     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 effects in descending order
sort(table(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
##     Accumulation     Intoxication     Biochemistry          Cell(s)
##               12               12               11                9
##       Physiology        Histology       Hormone(s)
##                7                5                1
```

Answer: The most common effect is population. The reason might be that the population of pollinators (like bees) are essential for biodiversity, agriculture, and food security.

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...]

```
# Get summary of the common Name column
summary(Neonics$Species.Common.Name)
```

```
##                          Honey Bee                      Parasitic Wasp
##                                667                                285
##              Buff Tailed Bumblebee                Carniolan Honey Bee
##                                183                                152
##                         Bumble Bee                   Italian Honeybee
##                                140                                113
##                    Japanese Beetle                  Asian Lady Beetle
```

```
##                              94                                  76
##                  Euonymus Scale                      Wireworm
##                              75                                  69
##               European Dark Bee                 Minute Pirate Bug
##                              66                                  62
##             Asian Citrus Psyllid                   Parastic Wasp
##                              60                                  58
##           Colorado Potato Beetle                  Parasitoid Wasp
##                              57                                  51
##             Erythrina Gall Wasp                    Beetle Order
##                              49                                  47
##       Snout Beetle Family, Weevil          Sevenspotted Lady Beetle
##                              47                                  46
##                 True Bug Order              Buff-tailed Bumblebee
##                              45                                  39
##                  Aphid Family                  Cabbage Looper
##                              38                                  38
##             Sweetpotato Whitefly                  Braconid Wasp
##                              37                                  33
##                 Cotton Aphid                  Predatory Mite
##                              33                                  33
##           Ladybird Beetle Family                     Parasitoid
##                              30                                  30
##                  Scarab Beetle                  Spring Tiphia
##                              29                                  29
##                  Thrip Order                Ground Beetle Family
##                              29                                  27
##               Rove Beetle Family                  Tobacco Aphid
##                              27                                  27
##                  Chalcid Wasp              Convergent Lady Beetle
##                              25                                  25
##                 Stingless Bee                  Spider/Mite Class
##                              25                                  24
##             Tobacco Flea Beetle                 Citrus Leafminer
##                              24                                  23
##                 Ladybird Beetle                       Mason Bee
##                              23                                  22
##                     Mosquito                    Argentine Ant
##                              22                                  21
##                      Beetle          Flatheaded Appletree Borer
##                              21                                  20
##             Horned Oak Gall Wasp                 Leaf Beetle Family
##                              20                                  20
##               Potato Leafhopper          Tooth-necked Fungus Beetle
##                              20                                  20
##                 Codling Moth           Black-spotted Lady Beetle
##                              19                                  18
##                 Calico Scale                 Fairyfly Parasitoid
##                              18                                  18
##                  Lady Beetle              Minute Parasitic Wasps
##                              18                                  18
##                    Mirid Bug                   Mulberry Pyralid
##                              18                                  18
##                     Silkworm                    Vedalia Beetle
```

```
##                                 18                                 18
##              Araneoid Spider Order                        Bee Order
##                                 17                                 17
##                     Egg Parasitoid                     Insect Class
##                                 17                                 17
##           Moth And Butterfly Order      Oystershell Scale Parasitoid
##                                 17                                 17
## Hemlock Woolly Adelgid Lady Beetle             Hemlock Wooly Adelgid
##                                 16                                 16
##                               Mite                       Onion Thrip
##                                 16                                 16
##              Western Flower Thrips                     Corn Earworm
##                                 15                                 14
##                 Green Peach Aphid                        House Fly
##                                 14                                 14
##                          Ox Beetle                Red Scale Parasite
##                                 14                                 14
##                 Spined Soldier Bug             Armoured Scale Family
##                                 14                                 13
##                  Diamondback Moth                    Eulophid Wasp
##                                 13                                 13
##                 Monarch Butterfly                    Predatory Bug
##                                 13                                 13
##              Yellow Fever Mosquito              Braconid Parasitoid
##                                 13                                 12
##                       Common Thrip      Eastern Subterranean Termite
##                                 12                                 12
##                             Jassid                       Mite Order
##                                 12                                 12
##                          Pea Aphid                  Pond Wolf Spider
##                                 12                                 12
##           Spotless Ladybird Beetle            Glasshouse Potato Wasp
##                                 11                                 10
##                           Lacewing           Southern House Mosquito
##                                 10                                 10
##            Two Spotted Lady Beetle                       Ant Family
##                                 10                                  9
##                       Apple Maggot                         (Other)
##                                  9                                670
```

```
# Sort by frequency
sort(table(Neonics$Species.Common.Name), decreasing = TRUE)[1:6]
```

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

Answer: The most frequently studied species are primarily pollinators or natural pest controllers, making them ecologically and economically vital. Their sensitivity to neonicotinoid pesticides is a major concern,as these species helps inform pesticide regulations, conservation efforts, and sustainable agricultural practices to ensure long-term environmental balance.

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...]

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

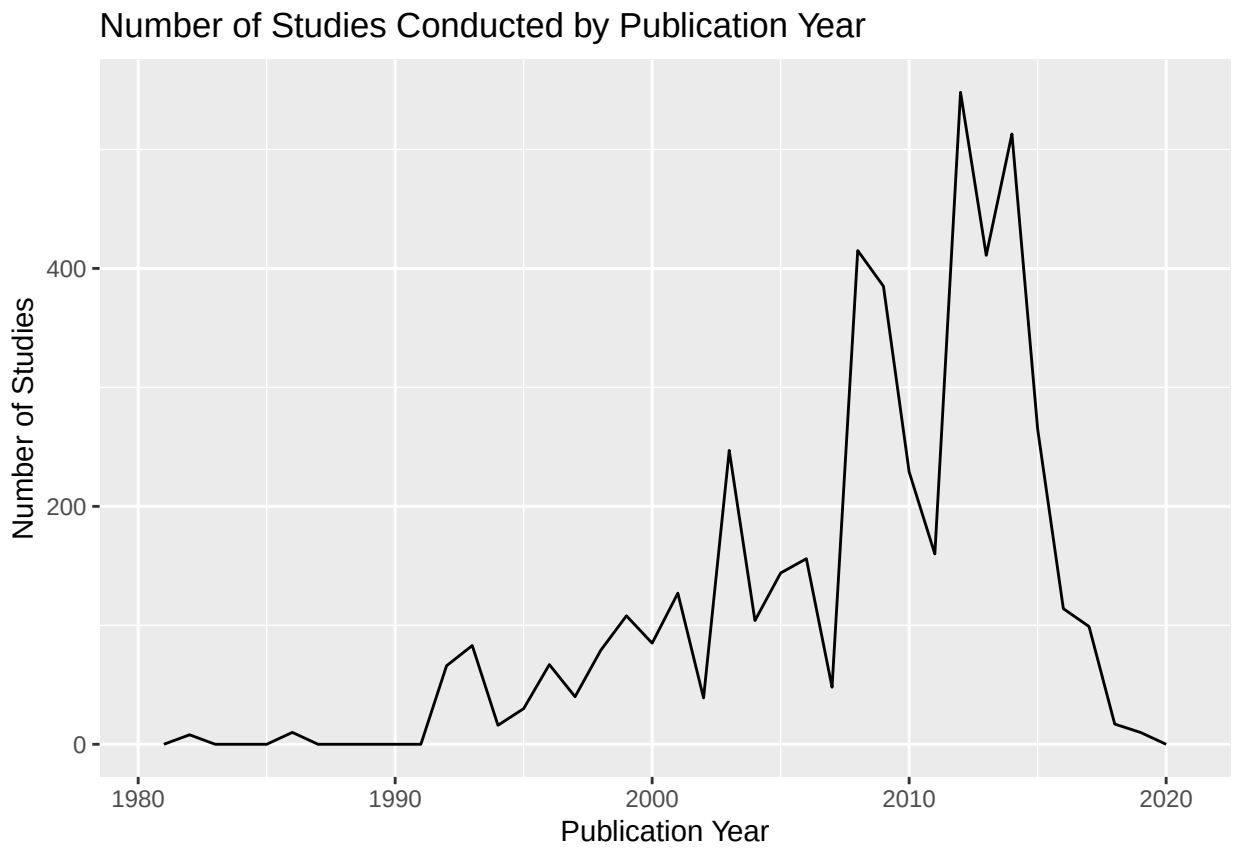
```
## [1] "factor"
```

Answer: The class of `Conc.1..Author.` column is factor, since the column has some text and inconsistent formatting (e.g., NR/,NR, 95.8,<0.5...)

Explore your data graphically (Neonics)

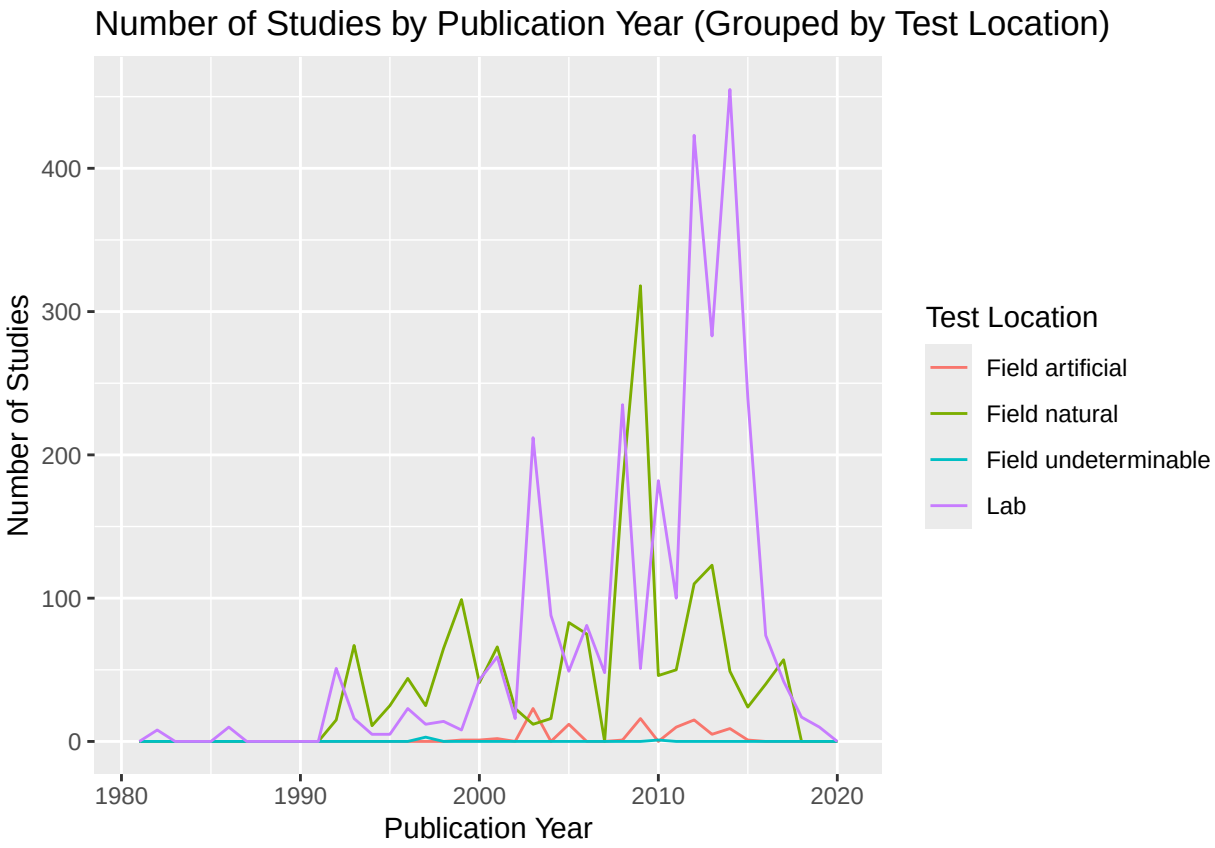
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
# Ensure "Publication.Year" column is Numeric
Neonics$Publication.Year <- as.numeric(as.character(Neonics$Publication.Year))


# Create the plot
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), binwidth = 1) +
  labs(title = "Number of Studies Conducted by Publication Year",
       x = "Publication Year",
       y = "Number of Studies")
```



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

```
# Create the frequency polygon plot with Test.Location as color
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), binwidth = 1) +
  labs(title = "Number of Studies by Publication Year (Grouped by Test Location)",
       x = "Publication Year",
       y = "Number of Studies",
       color = "Test Location")
```



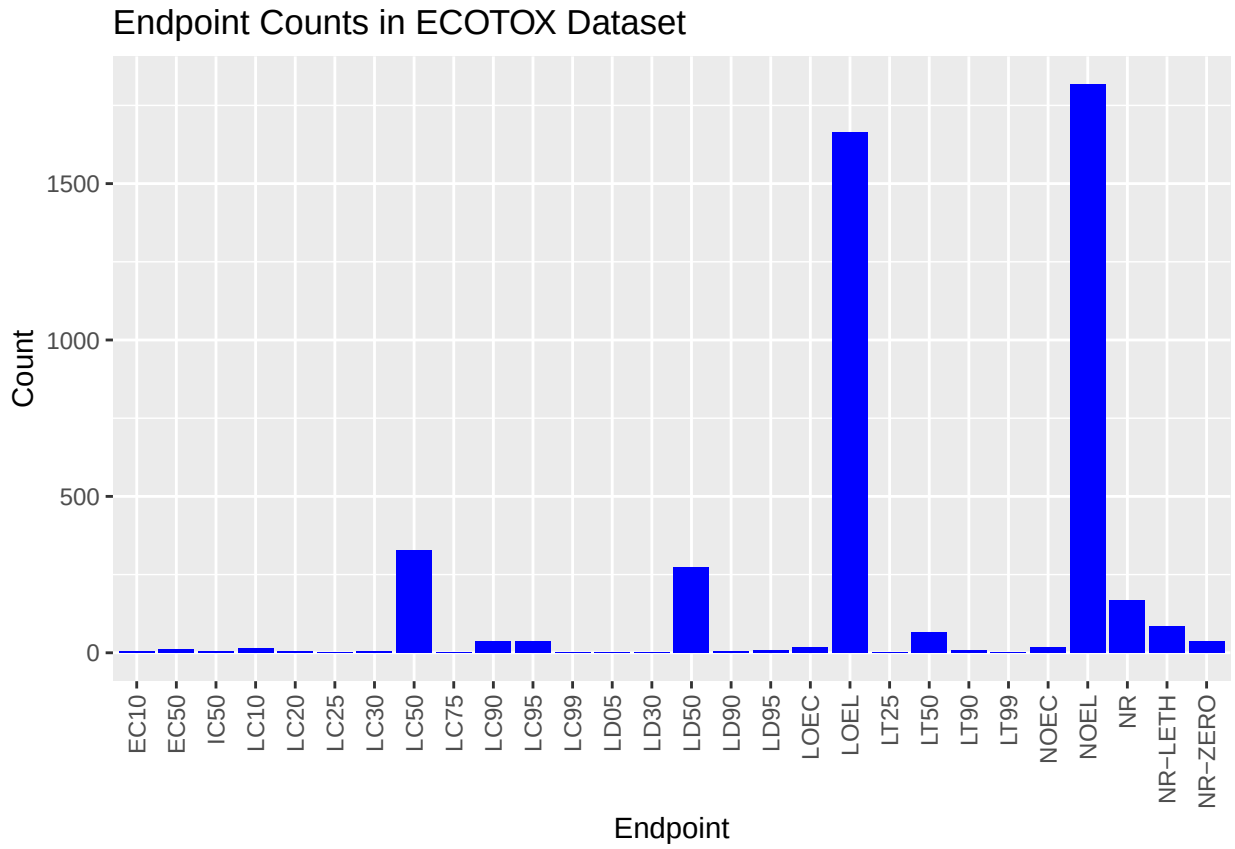
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations differ over time. The lab is generally the most common test location, except from 2008 to 2010 and from 1993 to 2000, when the field natural was the most common location.

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...]

```
# Create the bar plot for Endpoint counts
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar(fill = "blue") +
  labs(title = "Endpoint Counts in ECOTOX Dataset",
       x = "Endpoint",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



Answer: Two most common end points are LOEL and NOEL. LOEL indicates the lowest level at which harmful effects are observed; NOEL represents the highest level at which no harmful effects are seen.

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.

```
# Load the lubridate package
library(lubridate)

# Check the class of collectDate
class(Litter$collectDate)
```

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

```r
# Convert collectDate to a Date format if it's not
Litter$collectDate <- ymd(Litter$collectDate)

# Confirm the new class
class(Litter$collectDate)
```

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

```r
# Filter for unique dates in August 2018
unique(Litter$collectDate[Litter$collectDate >= ymd("2018-08-01") & Litter$collectDate <= ymd("2018-08-3
```

```
## [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`?

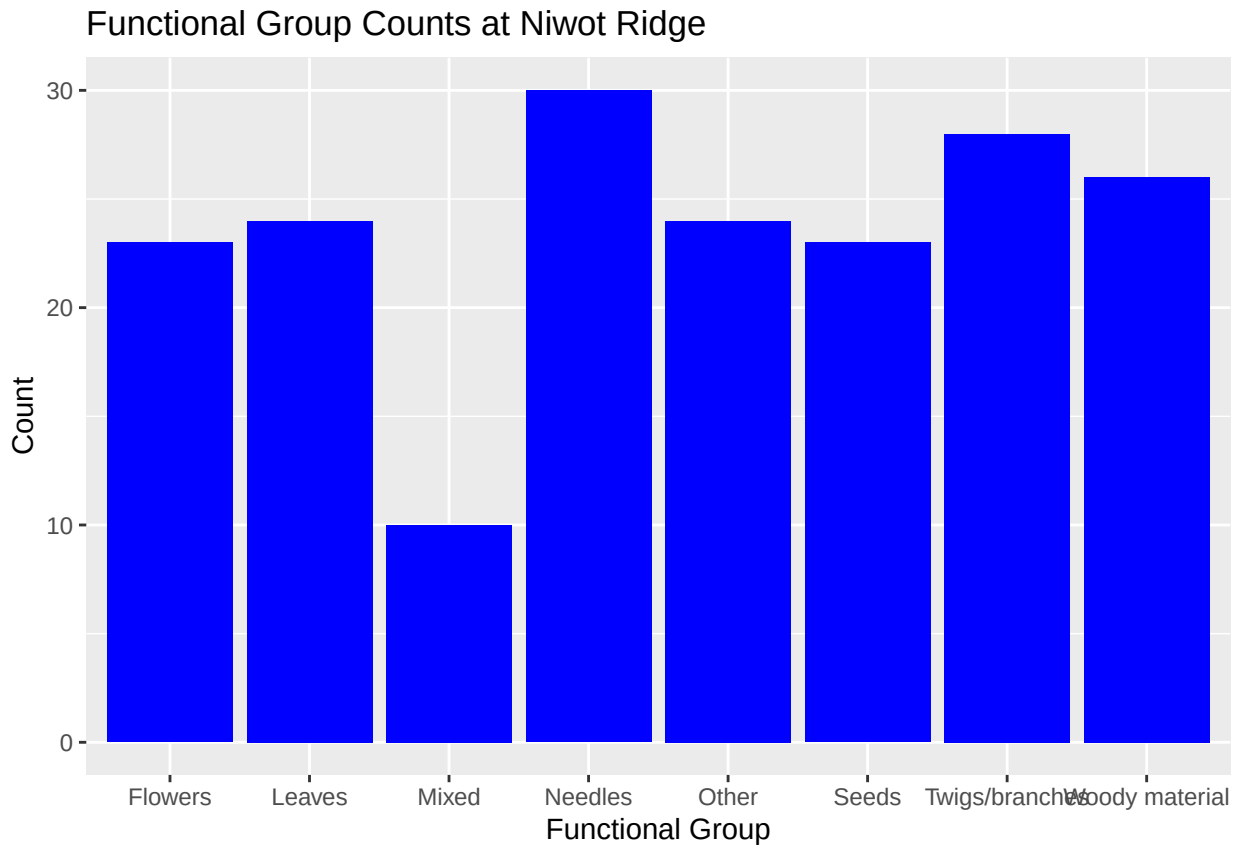
```r
# Get unique plot IDs
unique_plots <- unique(Litter$plotID)

# Count the number of unique plots
length(unique_plots)
```

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

Answer: Unique function can extract the distinct values and help identify unique entries without summarizing their distribution. But summary function provides an overview of data distribution (e.g., min, max, mean, median. . .)

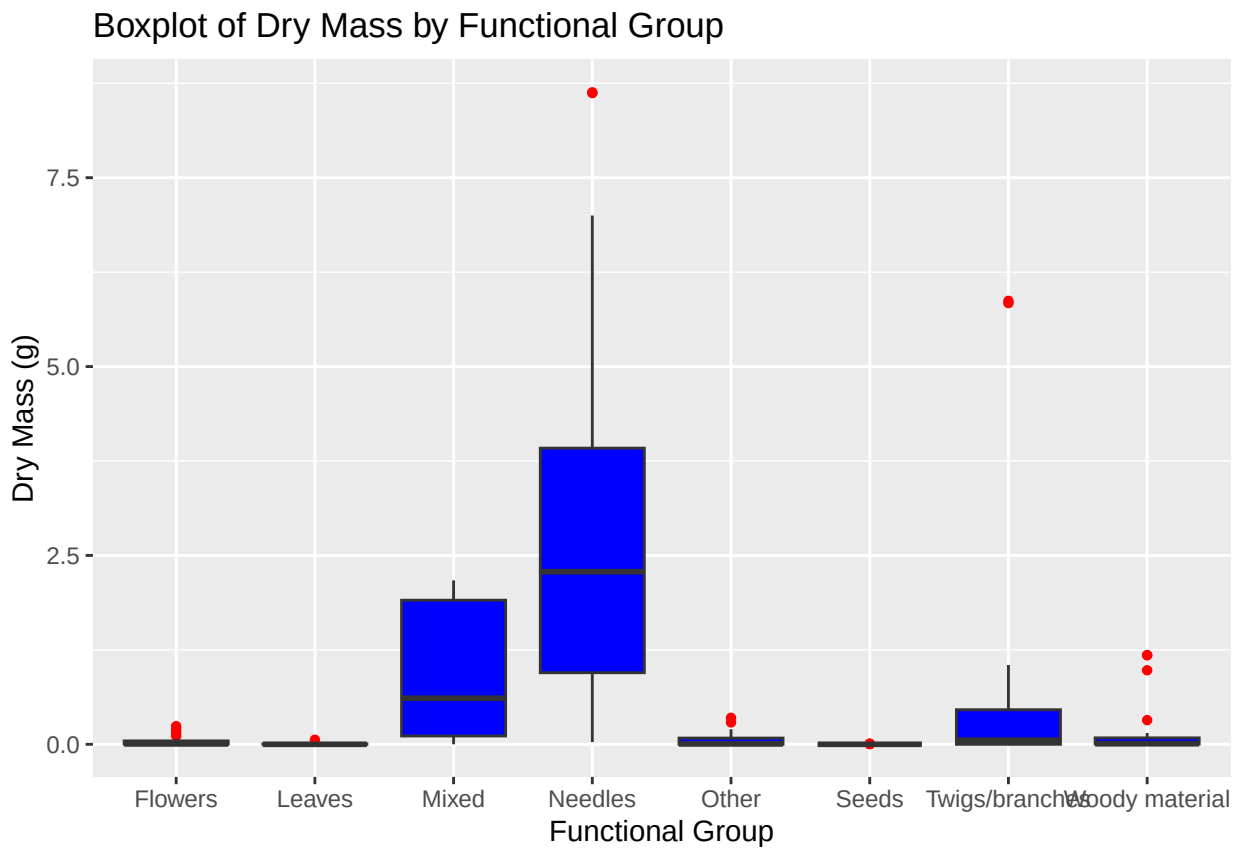
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
# Load required library
library(ggplot2)

# Create bar plot of functionalGroup counts
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar(fill = "blue") +
  labs(title = "Functional Group Counts at Niwot Ridge",
       x = "Functional Group",
       y = "Count")
```

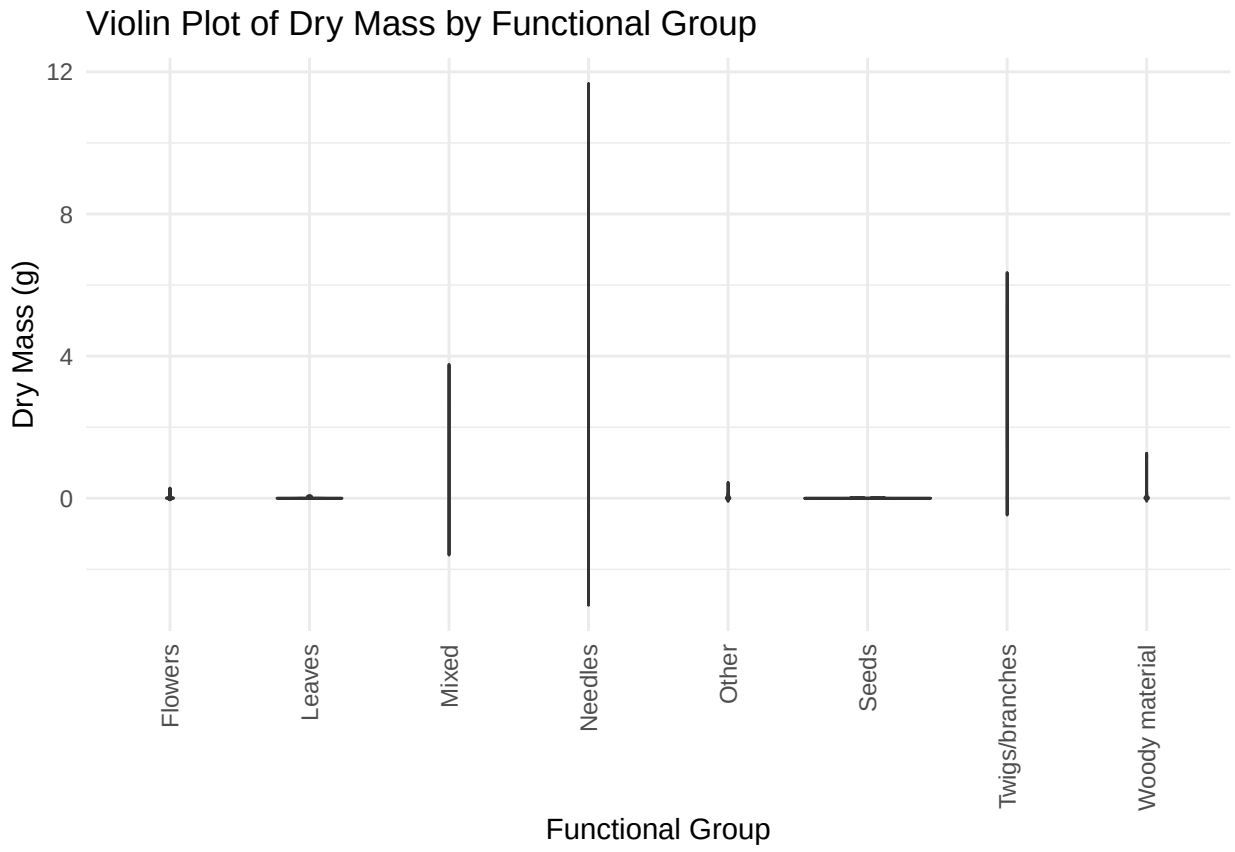


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
# Load required library
library(ggplot2)

# Boxplot of dryMass by functionalGroup
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_boxplot(fill = "blue", outlier.color = "red", outlier.shape = 16) +
  labs(title = "Boxplot of Dry Mass by Functional Group",
       x = "Functional Group",
       y = "Dry Mass (g)")
```




```
# Violin plot of dryMass by functionalGroup
ggplot(Litter, aes(x = functionalGroup, y = dryMass, fill = functionalGroup)) +
  geom_violin(trim = FALSE, alpha = 0.7) +
  theme_minimal() +
  labs(title = "Violin Plot of Dry Mass by Functional Group",
       x = "Functional Group",
       y = "Dry Mass (g)") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        legend.position = "none")
```



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

> Answer: The boxplot is more effective because it clearly shows summary statistics (median, quartiles, outliers) and avoids misleading density estimates for small sample sizes.

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

> Answer: Needles litter has the highest biomass.