Assignment 3: Data Exploration

Prisha

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
#Loading Packages
library(tidyverse); library(lubridate); library(here)

#Checking workspace
here()
```

```
## [1] "/Users/prisha/Desktop/EDE/EDE 2025"
```

```r
#Importing data
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.

   Answer: We can be interested in the ecotoxicology of neonicotinoids on insects for many reasons, one of them being that insects play key roles in ecosystems as pollinators, decomposers, and as a food source for other wildlife. The widespread use of neonicotinoids can disrupt these ecological roles, potentially leading to biodiversity loss and altered ecosystem functions.

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: Studying litter and woody debris in forests is important because they play key roles in nutrient cycling, carbon storage, and providing habitats for diverse organisms. They also influence soil health, forest regeneration, and fire dynamics, making them critical for understanding 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. Trap Types: Litter is collected using elevated PVC traps (0.5 m^2) and ground traps (3 m x 0.5 m) for larger woody debris 2. Sampling Frequency: Ground traps are sampled annually, while elevated traps are sampled biweekly in deciduous forests and every 1-2 months in evergreen sites. 3. Trap Placement: Traps are placed in tower plots, either randomly in dense vegetation or targeted in patchy areas, with one trap pair per 400 m^2.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
#Data dimensions
dim(Neonics)
```

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

```
colnames(Neonics)
```

```
##  [1] "CAS.Number"                       "Chemical.Name"
##  [3] "Chemical.Grade"                   "Chemical.Analysis.Method"
##  [5] "Chemical.Purity"                  "Species.Scientific.Name"
##  [7] "Species.Common.Name"              "Species.Group"
##  [9] "Organism.Lifestage"               "Organism.Age"
## [11] "Organism.Age.Units"               "Exposure.Type"
## [13] "Media.Type"                       "Test.Location"
## [15] "Number.of.Doses"                  "Conc.1.Type..Author."
## [17] "Conc.1..Author."                  "Conc.1.Units..Author."
## [19] "Effect"                           "Effect.Measurement"
## [21] "Endpoint"                         "Response.Site"
## [23] "Observed.Duration..Days."         "Observed.Duration.Units..Days."
## [25] "Author"                           "Reference.Number"
## [27] "Title"                            "Source"
## [29] "Publication.Year"                 "Summary.of.Additional.Parameters"
```

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 Effects
SummaryOfEffect <- summary(Neonics$Effect)
sort(SummaryOfEffect, decreasing = T)
```

```
##       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 effects that are studied include Population, Mortality, Behaviour, Feeding behaviour, and Reproduction. These effects are key as they impact insect survival, ecosystem functions, and population dynamics. Changes in behavior, feeding, and reproduction can disrupt pollination, food webs, and long-term species sustainability.

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…]

```
#Common Species
SpeciesComName <- summary(Neonics$Species.Common.Name, maxsum = 7)
SpeciesComName
```

```
##             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 commonly studied species in the dataset are Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, and Bumblebee. These species are key pollinators and play vital roles in agriculture and ecosystem health. They are of interest because their decline can directly impact food production, biodiversity, and ecological stability.

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…]

```
#Determining Class
class(Neonics$Conc.1..Author.)
```

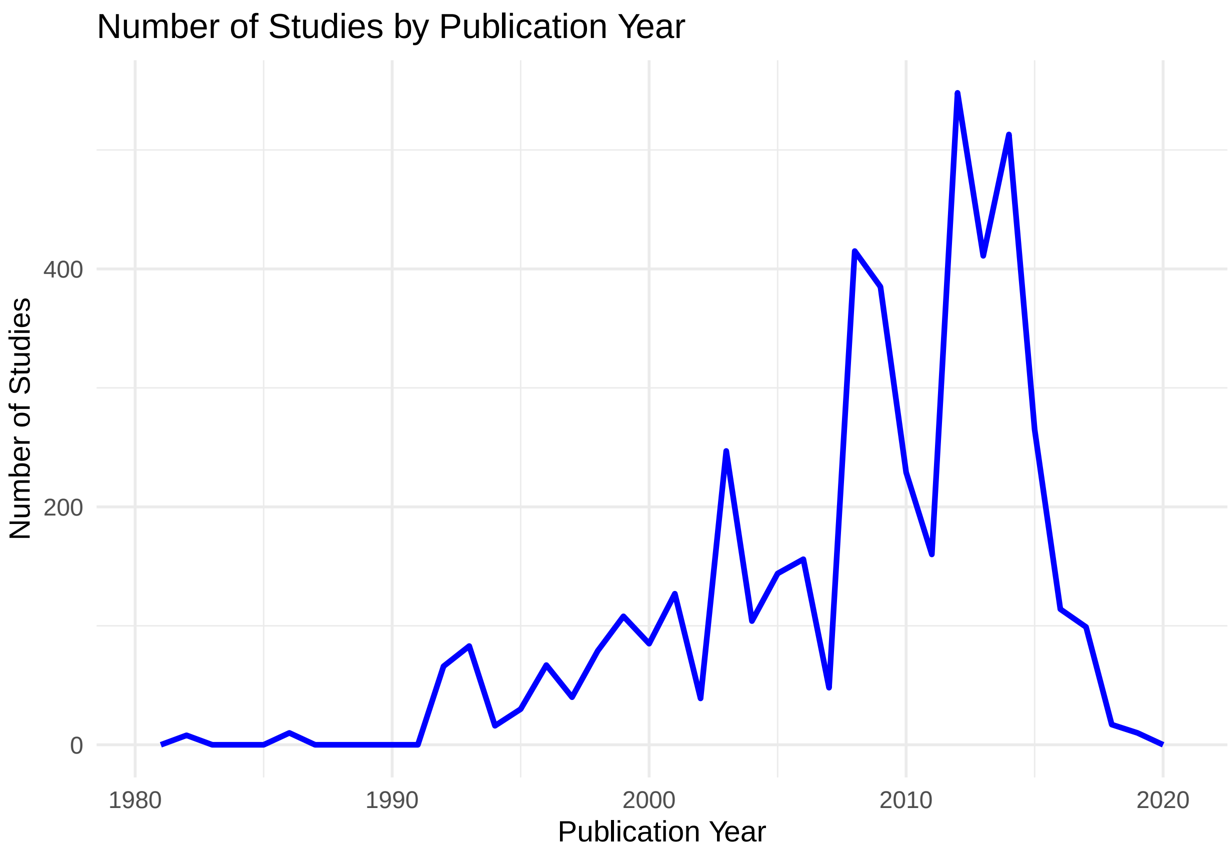
```
## [1] "factor"
```

Answer: The class of `Conc.1..Author.` is factor and not numeric because this column is a mix of numbers and text, and therefore R assumes these are categories rather than continuous data.

Explore your data graphically (Neonics)

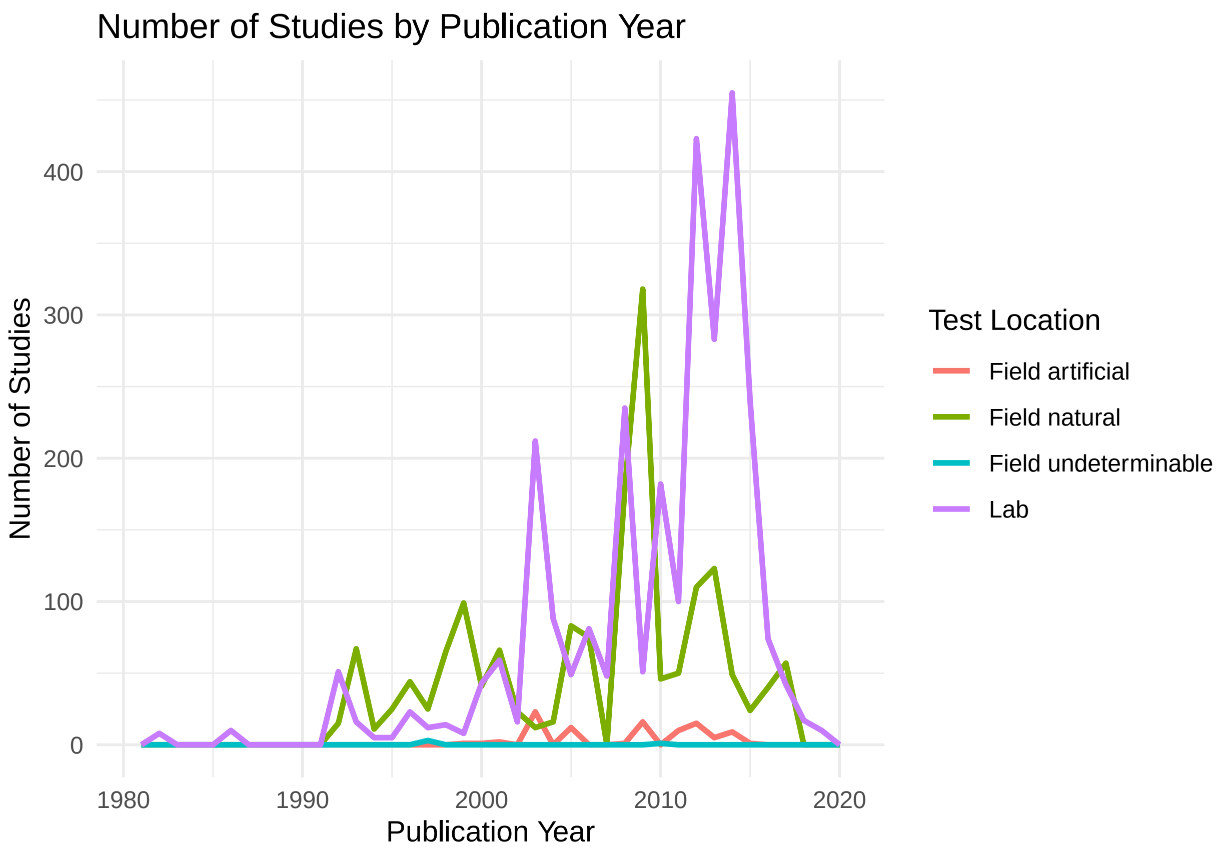
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
#Creating a frequency polygon graph
ggplot(Neonics, aes(x = Publication.Year)) +
  geom_freqpoly(binwidth = 1, color = "blue", linewidth = 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.

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



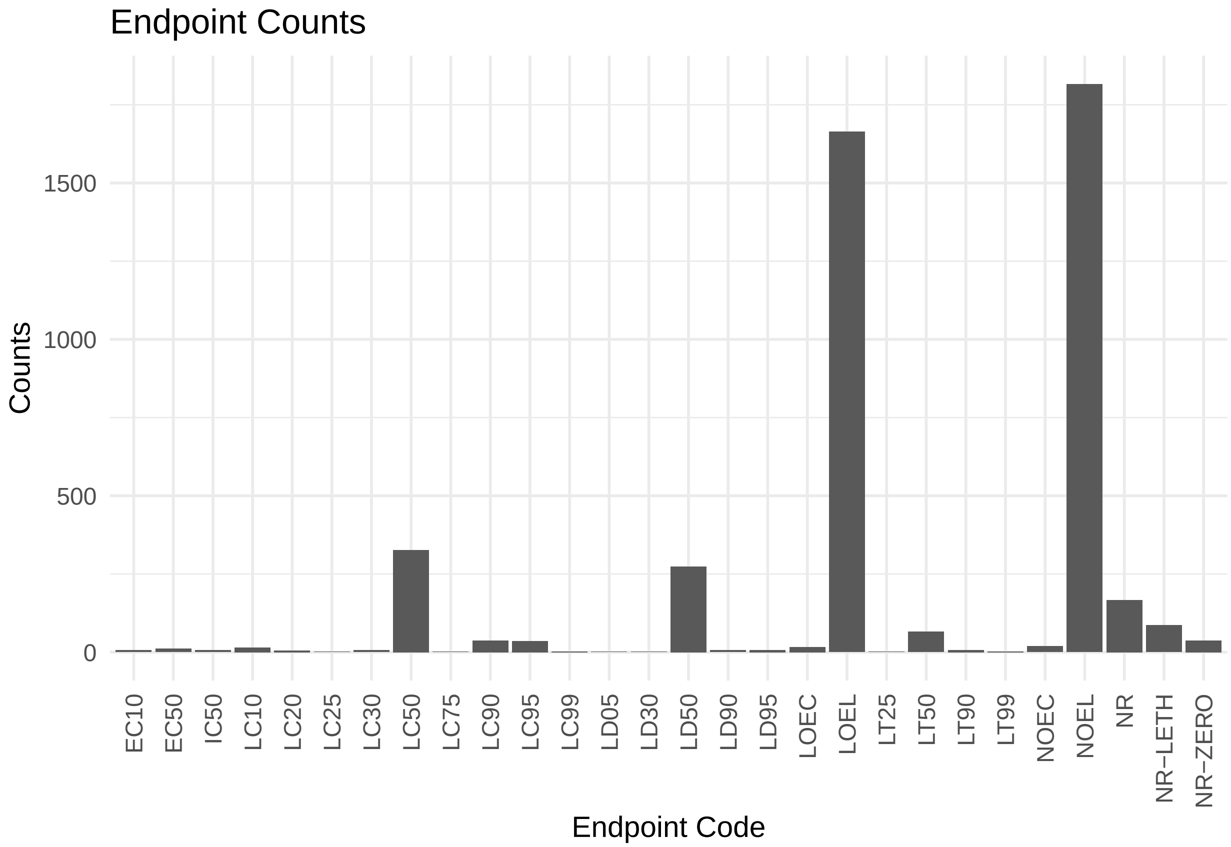
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The graph shows that lab studies were the most common test location, peaking between 2010–2015, followed by field natural studies, which also saw significant growth during this period. Field artificial and field undeterminable studies were rare throughout. After 2015, there's a noticeable decline in studies across all test locations.

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 graphs
ggplot(data = Neonics, aes(x = Endpoint)) +
  geom_bar() +
labs(title = "Endpoint Counts",
     x = "Endpoint Code",
     y = "Counts") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: The two most common endpoints are NOEL (No-Observable-Effect Level) and LOEL (Lowest-Observable-Effect Level). NOEL is defined as the highest dose or concentration that produces effects not significantly different from the control group, based on the statistical tests reported by the authors. In contrast, LOEL refers to the lowest dose or concentration that results in effects significantly different from the control group, as determined by the authors' statistical analyses.

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.

```
#Determining class
class(Litter$collectDate) #class = factor, not date
```

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

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

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

```r
unique_aug <- unique(Litter$collectDate[format(Litter$collectDate,"%Y-%m") == "2018-08"])
unique_aug
```

```
## [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
unique_plots <- unique(Litter$plotID)
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
```

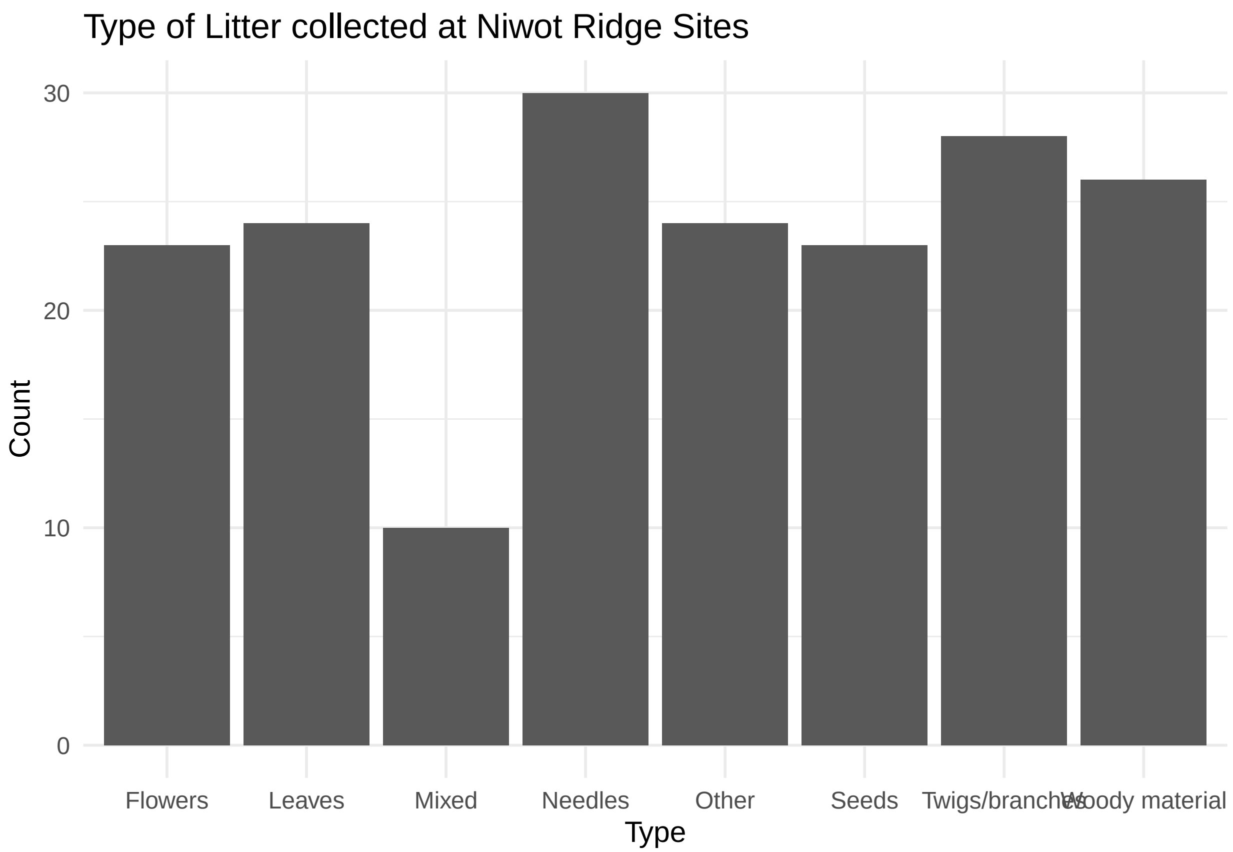
```r
summary(Litter$plotID)
```

```
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061
##       20       19       18       15       14        8       16       17
## NIWO_062 NIWO_063 NIWO_064 NIWO_067
##       14       14       16       17
```

Answer: Summary tells us how many times each plot ID was sampled, providing information on sampling frequency and basic statistics. In contrast, Unique identifies how many distinct plot IDs exist, giving the total number of different plots that were sampled.

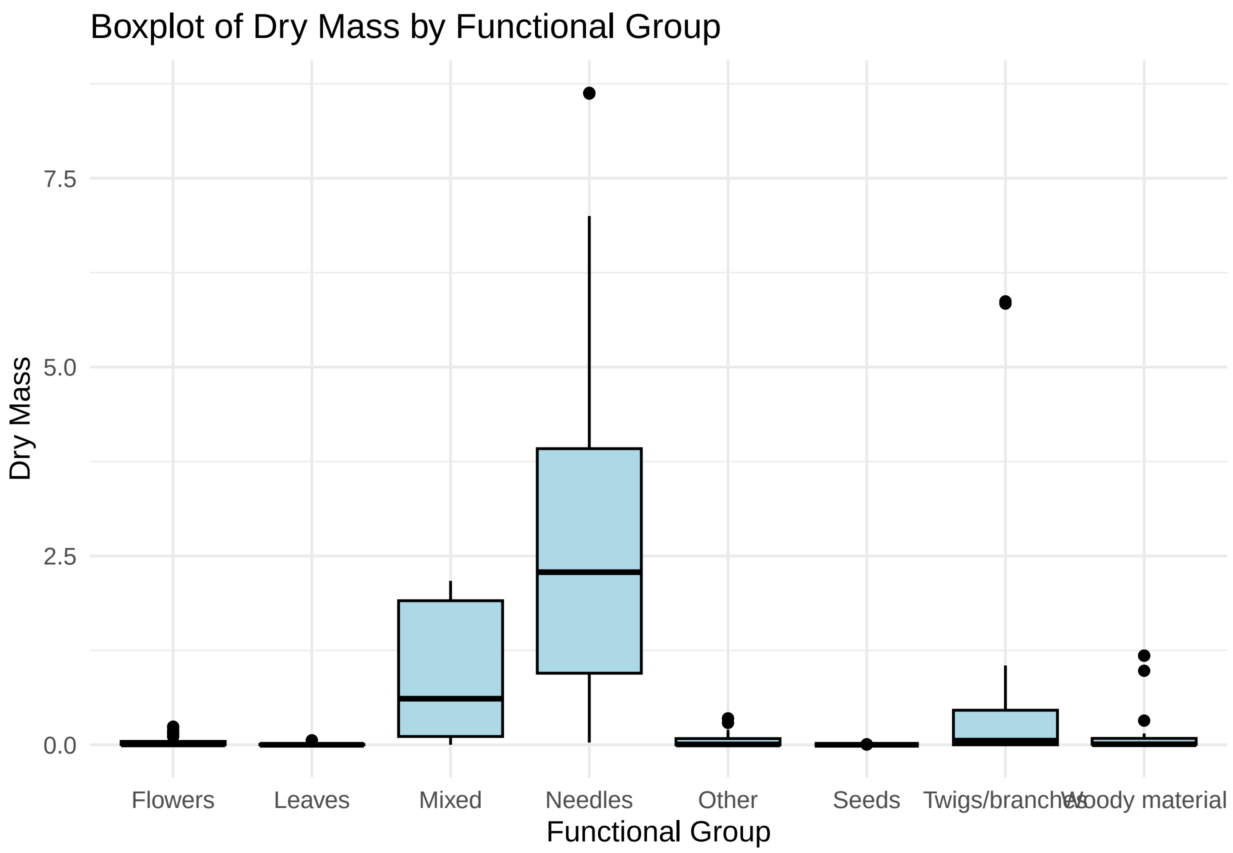
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
#Bar graphs
ggplot(data = Litter, aes(x = functionalGroup)) +
  geom_bar() +
labs(title = "Type of Litter collected at Niwot Ridge Sites",
     x = "Type",
     y = "Count") +
  theme_minimal()
```

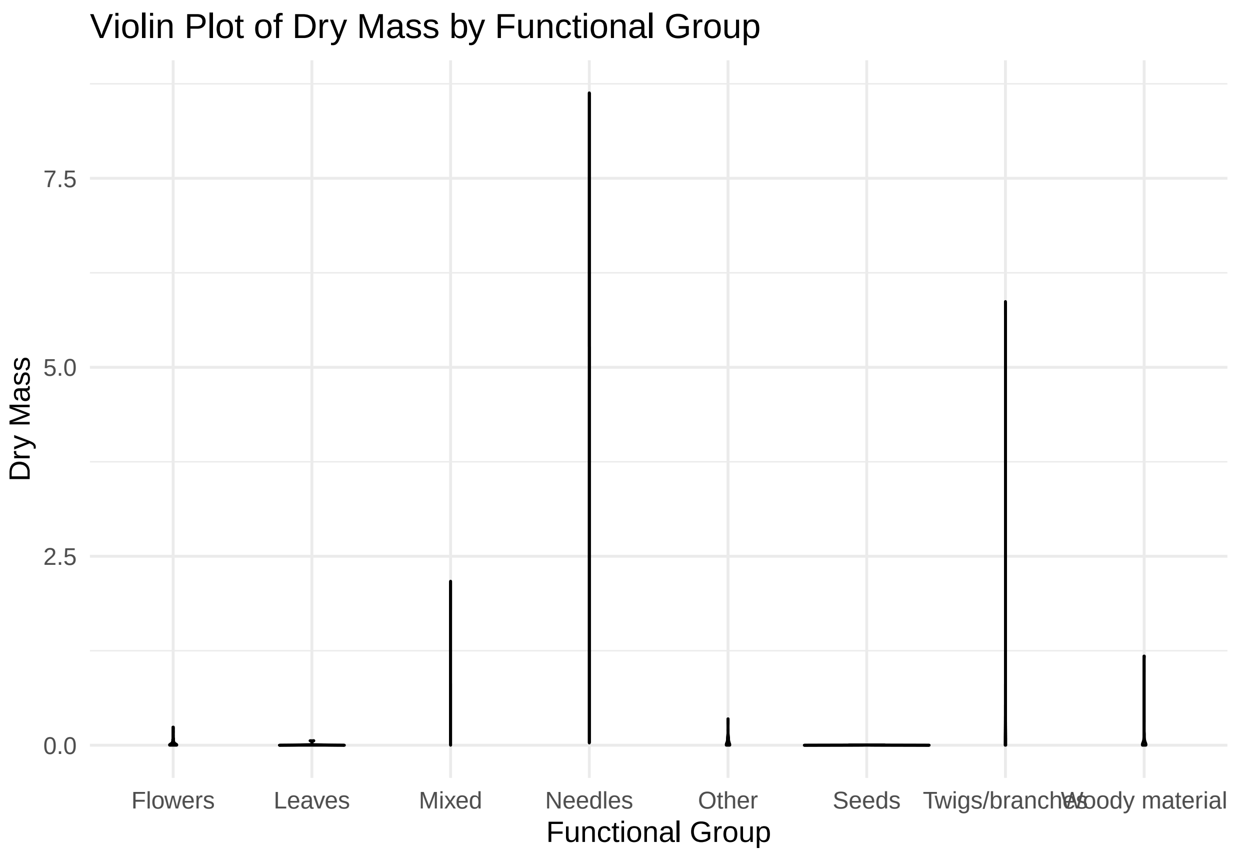


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#Boxplot and Violin plot
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Boxplot of Dry Mass by Functional Group",
       x = "Functional Group",
       y = "Dry Mass") +
  theme_minimal()
```



```
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_violin(fill = "lightgreen", color = "black") +
  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: The boxplot is more effective here because it provides clear, concise, and consistent summaries of the data, handles sparse data better, and visually highlights outliers, making it easier to compare functional groups.

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

> Answer: Based on the boxplot and violin plot, Needles tend to have the highest biomass at these sites. Mixed and Twigs/branches also show relatively higher biomass compared to other functional groups.