Assignment 3: Data Exploration

Tori Newton

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.

```r
#Load Packages
library(tidyverse)
library(lubridate)
library(here)
#Check current working directory
getwd()
```

```
## [1] "/home/guest/EDA_Spring2025"
```

```r
#Upload data sets
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 might be interested in the ecotoxicology of neonicotinoids on insects because it is important to know the impacts on pollinators for an agriculture insecticide. For example, if it is determined that these neonicotinoids negatively impact bees and other beneficial insects, then they should not be used in agriculture since the plants rely on pollinator species.

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 that falls on the ground in forests is important for understanding ecosystem processes, carbon cycling, and biodiversity. For example, decomposing litter and woody debris release essential nutrients like nitrogen back into the soil, supporting tree and plant growth. Decomposing material also provides habitat for insects, fungi, and amphibians. Analyzing litter and woody debris datasets also allow scientists to track how forests change over time and identify trends linked to climate change, pollution, or land use changes.

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. Use of elevated litter traps and ground traps for woody debris. These traps are strategically placed within designated plots to capture representative samples. 2. Temporal sampling design: ground traps are sampled once per year to ensure consistenet data over time. 3. After collection, debris is sorted into categories based on plant functional type to aid in data analysis.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
#determine dimensions of the dataset Neonics
dim(Neonics)
```

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

```r
#Neonics has 4623 rows and 30 columns
```

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...]

```r
#Using the 'summary' function on 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
```

```r
#sort(Neonics$Effect, decreasing = TRUE) #commented out before knitting to avoid
#displaying the long list
#Reproduction and population are the most common effects that are studied.
```

Answer:Reproduction and population are of particular interest in ecotoxicology because it is important to understand the long term ecological consequences of neonicotinoid insecticides. Neonicotinoids don't just kill insects, they weaken populations over generations by reducing their ability to reproduce and sustain population numbers. Insects are vital for pollination, food webs, and ecosystem balance so these effects are significant for environmental health and human agriculture.

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
#Using the 'summary' function on the "common name" column
summary(Neonics$Species.Common.Name, maxsum = 6)
```

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

```
#The six most commonly studied species in the dataset are the honey bee,
#parasitic wasp, buff tailed bumblebee, Carniolan honey bee, bumble bee,and
#other
```

Answer:These species are key pollinators in ecosystems and agriculture. They are critical for pollinating crops and pollinator decline due to neonicotinoids poses a serious threat to biodiversity and food security.

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…]

```
#View the dataset
#View(Neonics) #This was commented out in order to knit
#Determining class of 'Conc.1..Author'
class('Conc.1..Author')
```

```
## [1] "character"
```

```
#The class of 'Conc.1..Author' is character
```

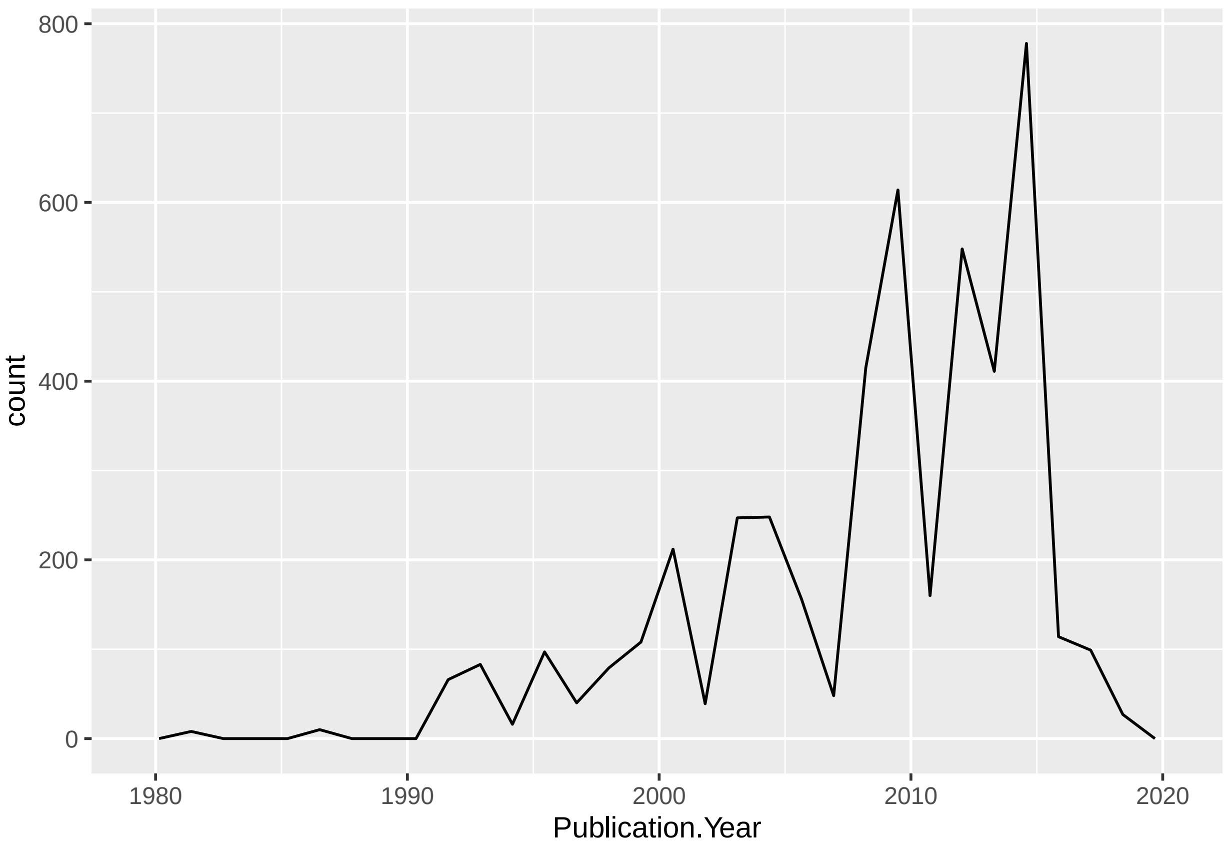
Answer:The class of ‘Conc.1..Author’ is a character, not numeric, because there are many missing values or not recorded data points present in the dataset. Since the column contains non-numeric values, R will treat it as a character.

Explore your data graphically (Neonics)

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

```
#Load ggplot2
library(ggplot2)
#Generating a plot using 'geom_freqpoly'
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year))
```

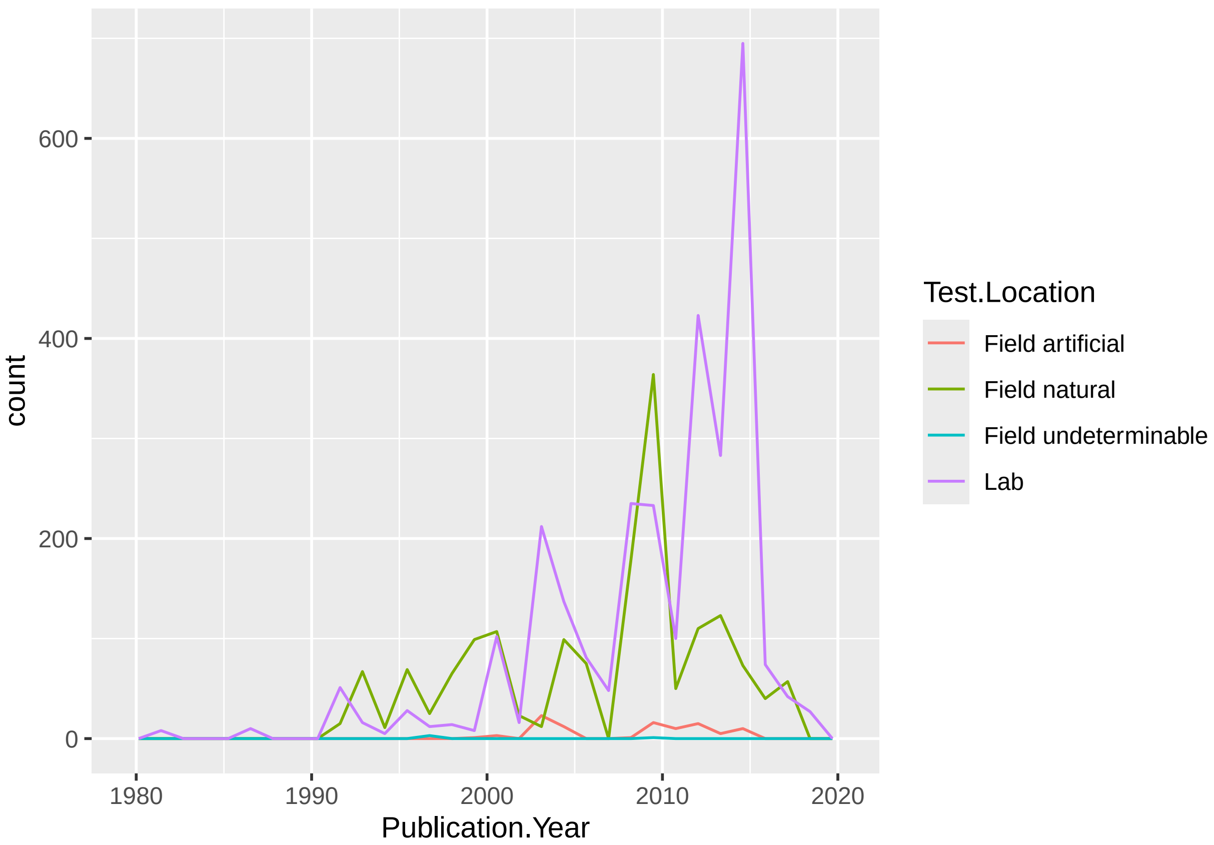
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



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

```
#Adding a color aesthetic
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, colour = Test.Location))
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



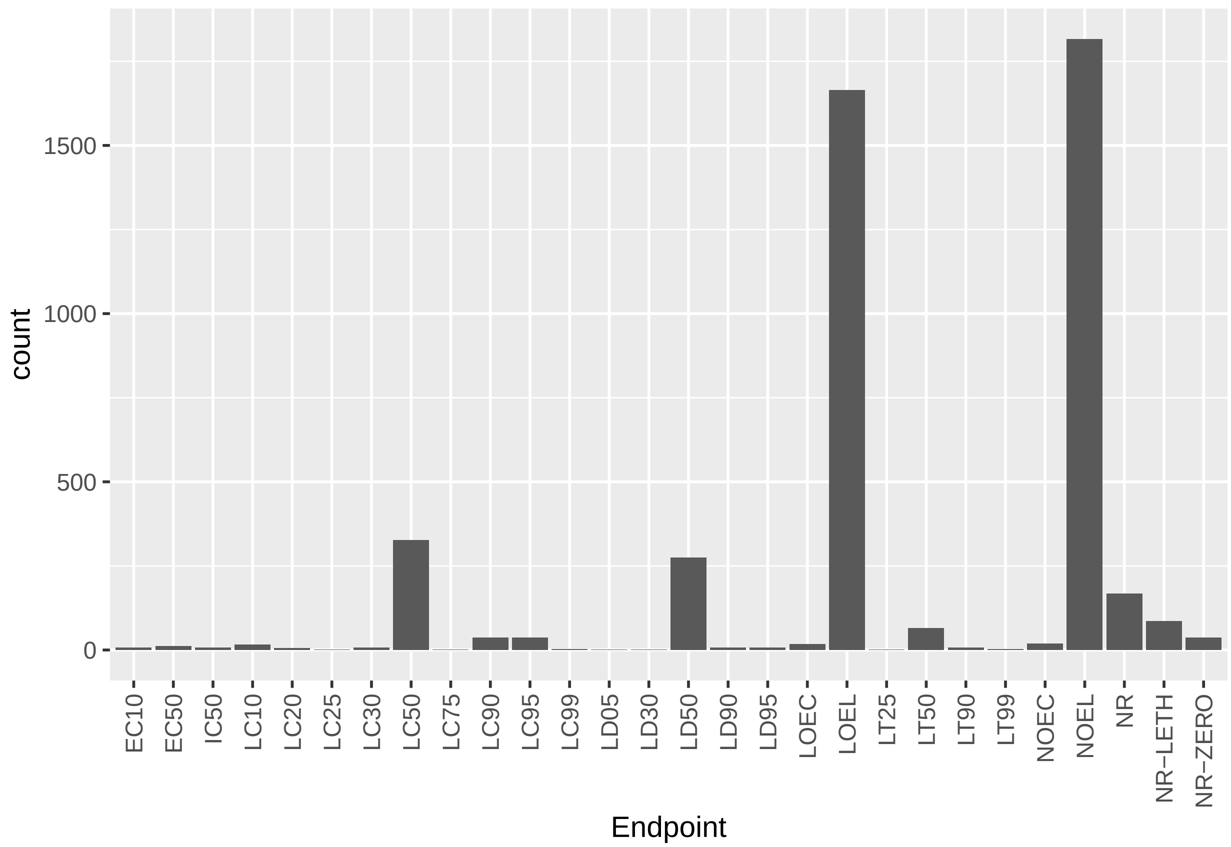
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: Overall, the most common test locations are in the lab. Natural field experiments were also popular, especially in 2010. Lab is typically the most common test location for all years, especially from 2010-2020. For the years of about 1992-2000, natural field experiments were more common than lab.

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…]

```
#Creating a 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))
```



Answer:The two most common end points are LOEL and NOEL. The endpoint 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.The endpoint LOEL means lowest-observable-effect-level: lowest does (concentration) producing effects that were significantly different from responses of controls.

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"
```

```
#collectDate is currently a factor
#Change collectDate to a date
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
class(Litter$collectDate)
```

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

```
#collectDate is now a date
#Extract the dates from August 2018
unique(Litter$collectDate[format(Litter$collectDate, "%Y-%m") == "2018-08"])
```

```
## [1] "2018-08-02" "2018-08-30"
```

```
#Litter was sampled on August 2 and 30 in 2018.
```

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`?

```
#Determine how many different plots were sampled at Niwot Ridge
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_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

```
length(unique(Litter$plotID))
```

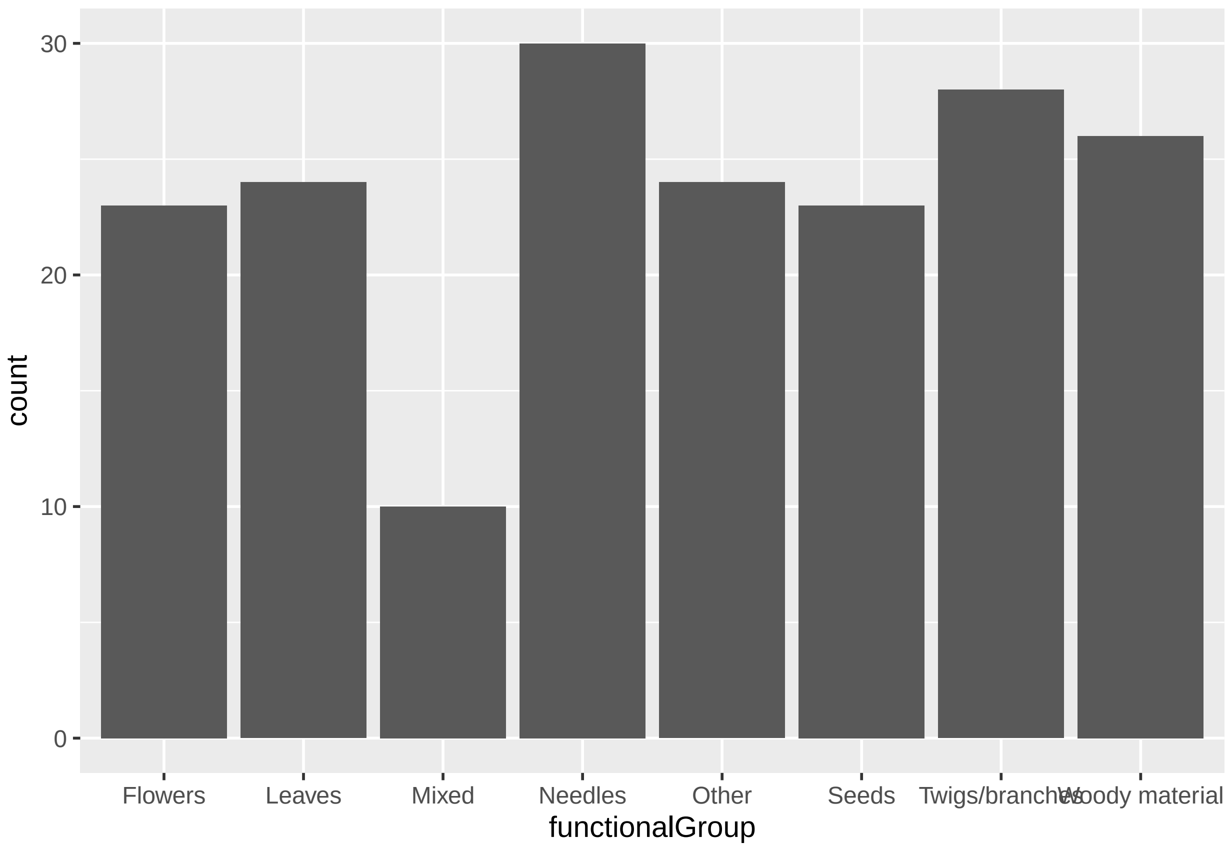
```
## [1] 12
```

```
#12 different plots were sampled at Niwot Ridge.
```

Answer: 'summary' provides an overall summary of the dataset while 'unique' returns a list of unique values in a specific column.

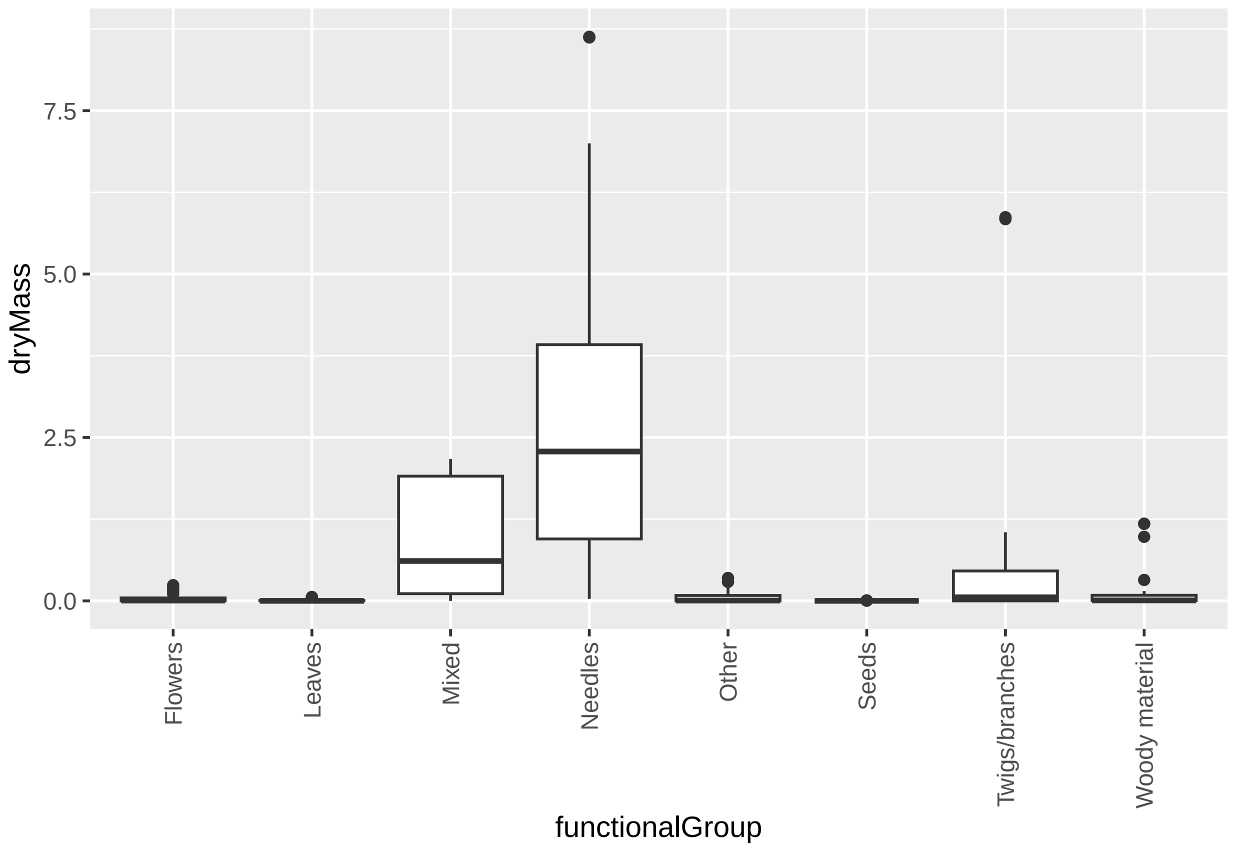
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#Create a bar graph of functionalGroup counts
ggplot(Litter, aes(x = functionalGroup,)) +
  geom_bar()
```

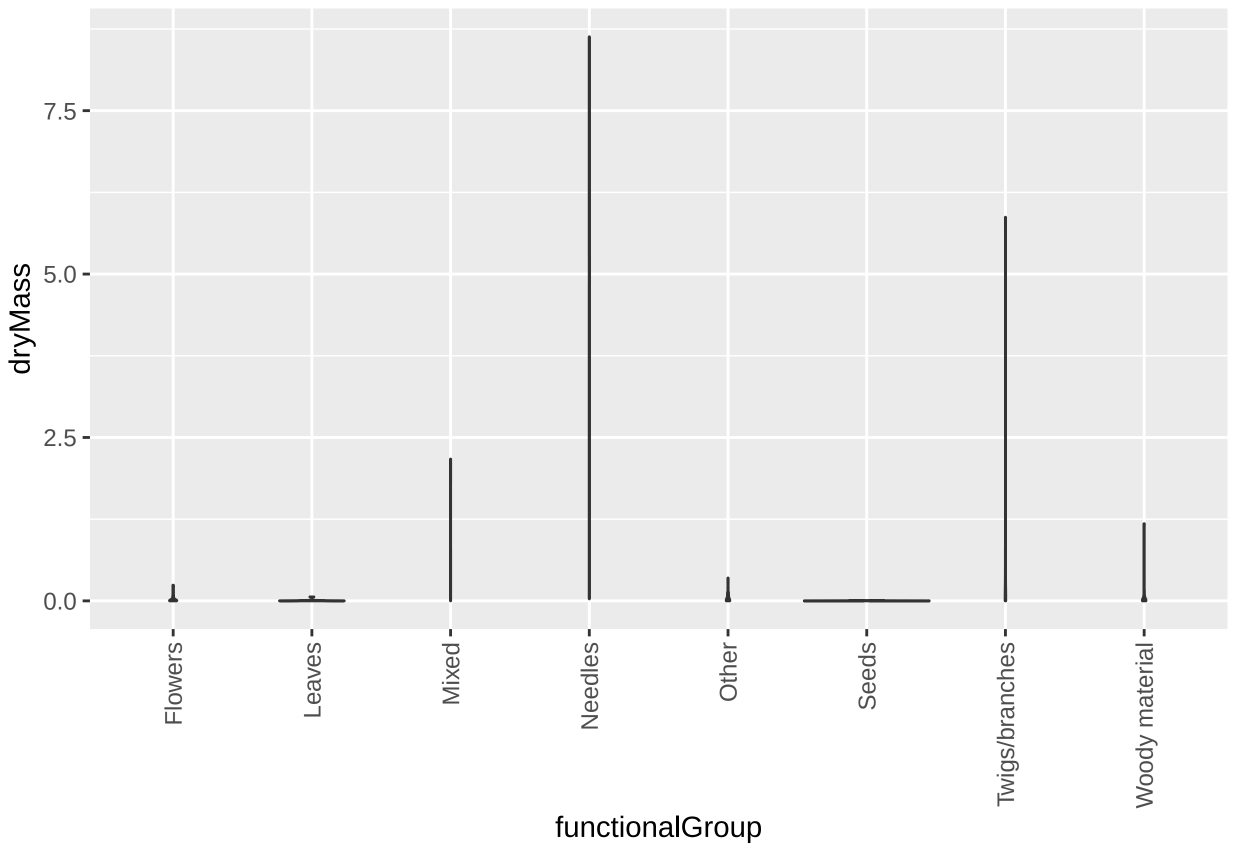


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#Create a boxplot of dryMass by functionalGroup
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) + geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



```
#Create a violin plot of dryMass by functionalGroup
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) + geom_violin() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



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

> Answer: Boxplots are useful for comparing distributions at a glance and shows a more clear comparison of summary statistics.

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

> Answer: Needles, Mixed, and Twigs/branches are the types of litter that tend to have the highest biomass at these sites.