Assignment 3: Data Exploration

David Liddle

Fall 2023

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 Sakai.

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. Check your working directory, load necessary packages (tidyverse, lubridate), 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.

```
#install.packages("tidyverse") #install tinyverse package
#install.packages("lubridate") #install lubridate package
library(tidyverse) #load tidyverse package
library(lubridate) #load lubridate package
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",
                    stringsAsFactors = TRUE) #Upload "Neonics" dataset
Litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv",
                   stringsAsFactors = TRUE) #Upload "Litter" dataset
```

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: According to foodandwaterwatch.org, Neonicotinoids are highly toxic to pollinating insects like bees. Bees play a critical role in the pollination of the crops we eat, and they are already in the middle of a large ecological crisis due to colony collapse disorder. Using these pesticides could further exacerbate the threats they face and may do more harm than good for our crop yield. Hence, the effect of neonicotinoids on pollinating insects should be closely monitored.

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 on the forest floor is important for many reasons. By measuring the amount of litter on the forest floor, we can measure the amount of litter input and heterotrophic decomposition, providing valuable information on carbon sequestration and nutrient cycling.

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 is conducted at terrestrial NEON sites that have woody vegetation taller than 2 meters. This sampling occurs exclusively in tower plots. The locations of these tower plots are chosen randomly within the 90% flux footprint of the primary and secondary airsheds. 2. In sites with forested tower airsheds, litter sampling is targeted to take place in 20 plots, each measuring 40 meters by 40 meters. In sites with low-statured vegetation over the tower airsheds, litter sampling is targeted to take place in 4 tower plots (40m x 40m, for co-located soil sampling) and 26 smaller plots, each measuring 20 meters by 20 meters. 3. Ground traps are sampled once per year. The target sampling frequency for elevated traps varies by vegetation present at the site, with sampling occurring once every two weeks in deciduous forests sites during senescence, and sampling once every 1-2 months year-round in evergreen sites.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) #Calculate the dimensions of the Neonics dataset
```

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

```
# The dimensions of the dataset are 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?

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

```
##     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
```

Answer: The most commonly studied effects are population, mortality, and behavior. These subjects are of specific interest because they directly correlate to the effectiveness of the pesticide on the insect. These research topics provide the most insight on whether a pesticide is working effectively on a target insect, and what the unforeseen impacts on non-target insects, and thus the ecosystem, may be.

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: The `sort()` command can sort the output of the summary command...]

```
#Generate a summary table of the studied insects in the dataset
Neonics_species_summary <- summary(Neonics$Species.Common.Name)
#Sort the summary table in order of most to least common
Neonics_species_summary_sorted <- sort(Neonics_species_summary, decreasing = TRUE)
#Generate only the top 6 most commonly studied species
head(Neonics_species_summary_sorted, n=6)
```

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

Answer: The first most commonly studied species is classified as 'Other'. However, the next five most studied are the Honey Bee, Parasitic Wasp, Buff Tailed Bumble Bee, Carniolan Honey Bee, and Bumble Bee. Aside from the wasp, the next most-studied insects are all bees, which are important pollinators. These studies are likely researching the unintended effects these pesticides have on pollinators.

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?

```r
#Identify the class of the "Conc.1..Author" column
class(Neonics$Conc.1..Author.)
```

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

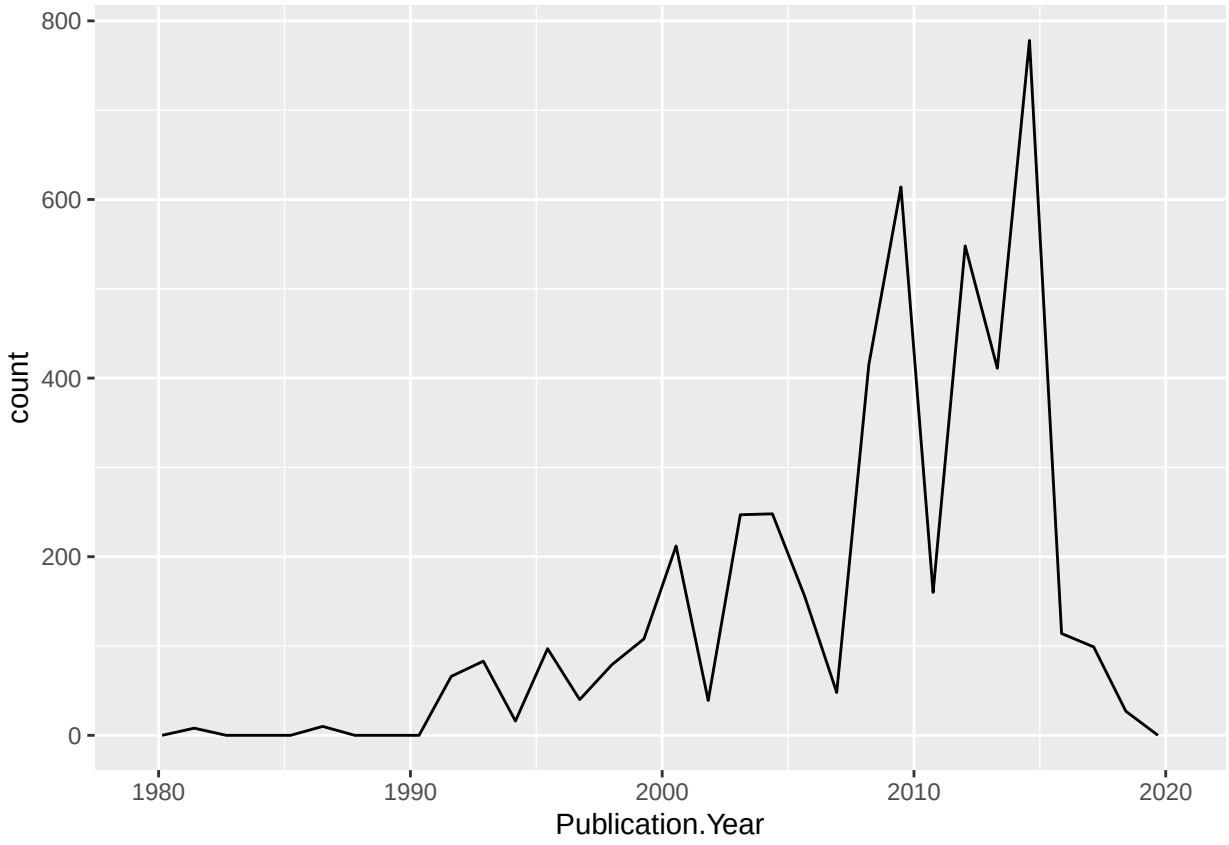
Answer: The column "Conc.1..Author" is a factor and not a numeric value, because when we imported the data set using the read csv command, we imported the dataset as a string of factors, which changed the classification of the numeric values.

Explore your data graphically (Neonics)

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

```r
library(ggplot2) # Load ggplot package
# Create linegraph of the number of publications by year.
Neonics_Linegraph <- ggplot(Neonics, aes(x = Publication.Year)) + geom_freqpoly()
Neonics_Linegraph #Print graph
```

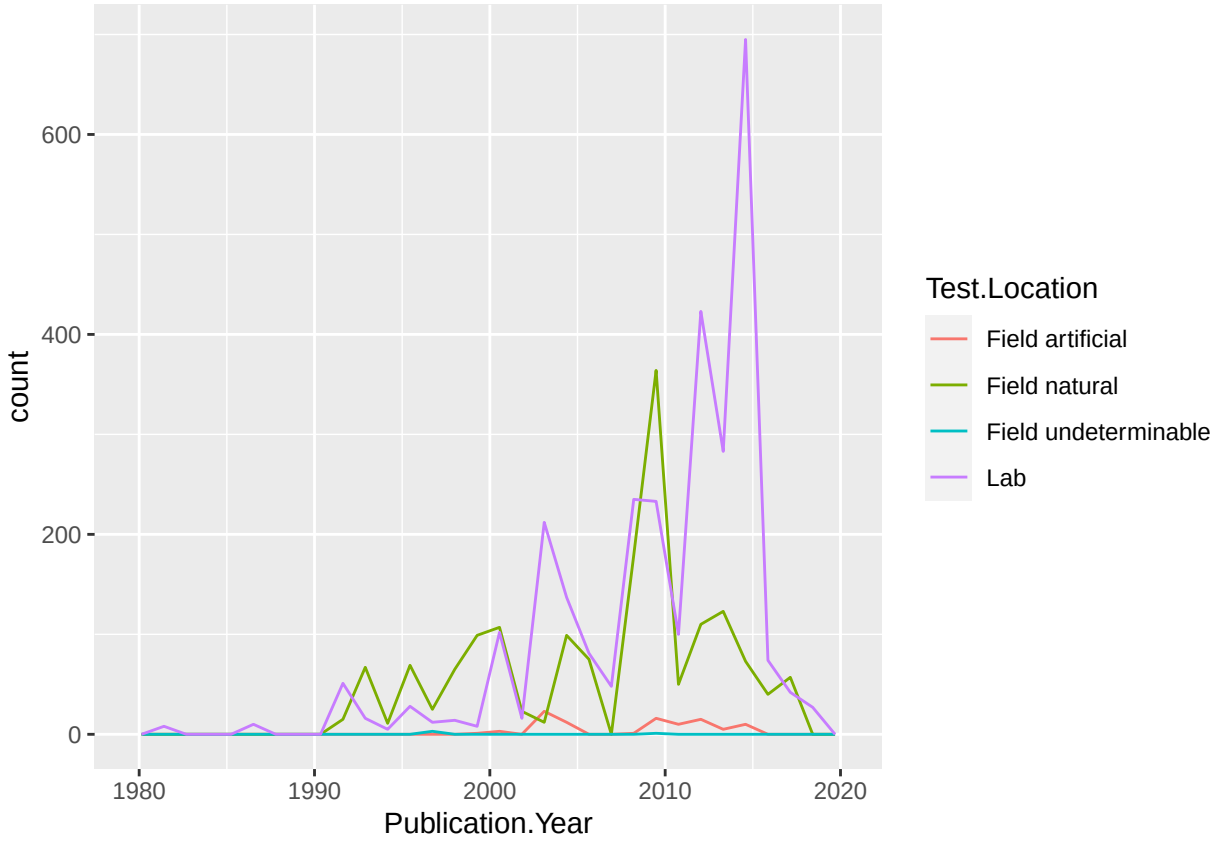
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
#Regenerate graph and add a color command
Neonics_Linegraph_Color <-ggplot(Neonics,
            aes(x = Publication.Year, color = Test.Location)) + geom_freqpoly()
Neonics_Linegraph_Color
```

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



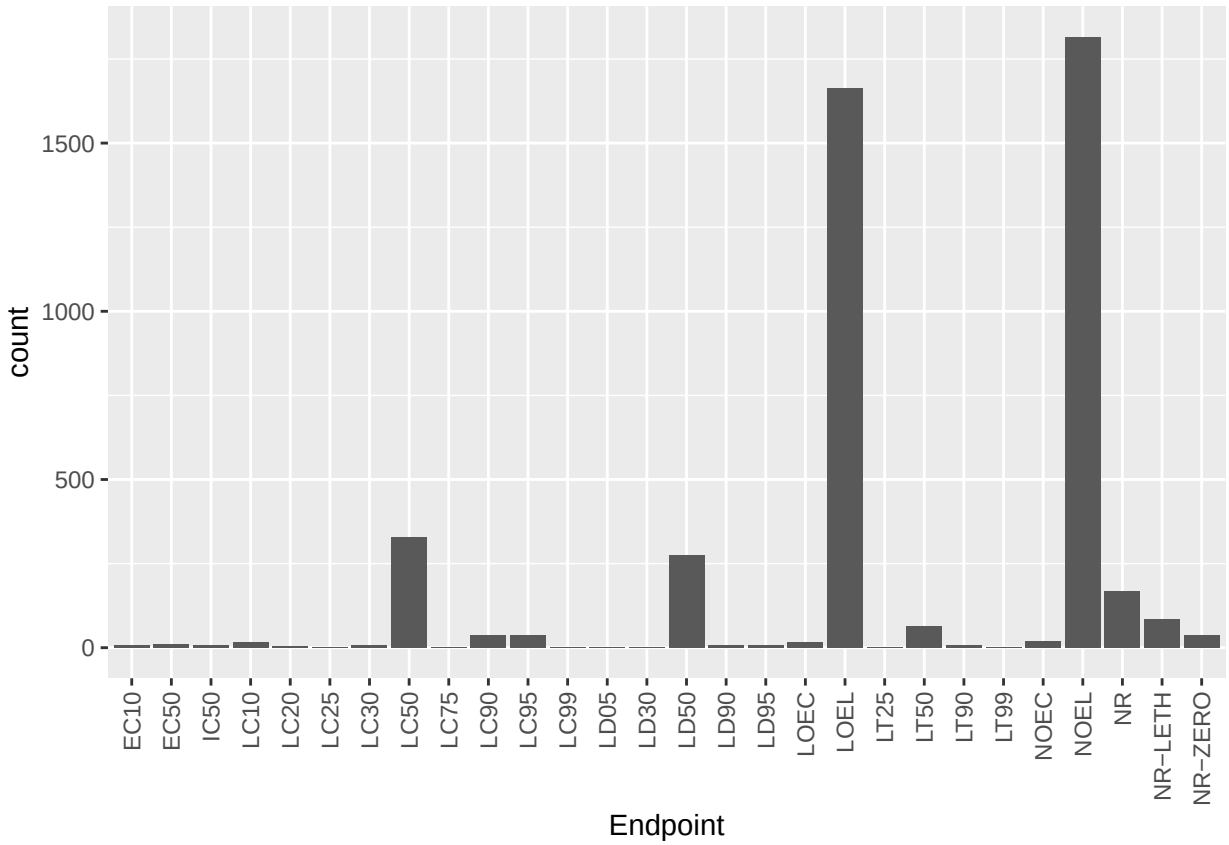
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test location is the lab. However, during the 1990s, the field natural location was more common until the early 2000s when lab tests skyrocketed and peaked around 2015.

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 bar graph with endpoint counts, with x-axis label adjustment
Neonics_bargraph <- ggplot(Neonics,
      aes(x =Endpoint)) + geom_bar() + theme(axis.text.x = element_text(angle = 90,
      vjust = 0.5, hjust=1))
Neonics_bargraph
```



Answer: The two most common endpoints are the NOEL and LOEL endpoints, respectively. LOEL stands for lowest-observable-effect-level, which is the lowest dose producing effects that were significantly different. NOEL is the no-observable-effect-level, which is the highest dose not producing significant changes from the control samples.

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.

```
class(Litter$collectDate) #Determine Class
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate) #Reformat as date
class(Litter$collectDate) #Confirm Class
```

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

```
#Extract the sampling dates from August 2018
august_dates <- unique(Litter$collectDate[month(Litter$collectDate) == 8])
august_dates
```

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

```
# The two sampled dates in August 2018 were 8/2/2018 and 8/30/2018.
```

13. Using the `unique` function, determine how many plots were sampled at Niwot Ridge. How is the information obtained from `unique` different from that obtained from `summary`?

```
summary(Litter$namedLocation) # show summary of plots sampled at Niwot Ridge
```

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

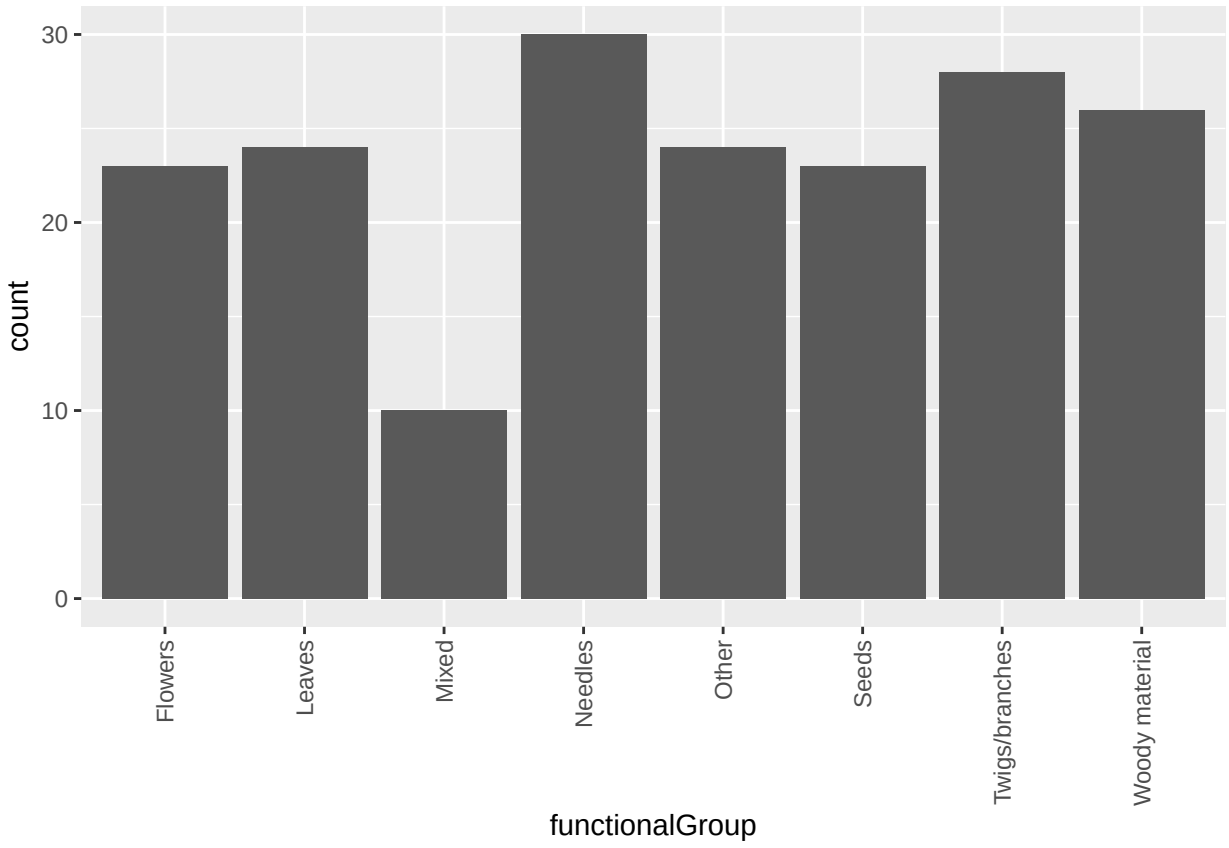
```
NIWO <- unique(Litter$namedLocation) #Extract the unique number of plots sampled at Niwot Ridge
length(NIWO) # find length
```

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

Answer: The summary function details the number of samples taken at each plot, whereas the unique function identifies the separate plot names within the namedLocation column. A total of 12 unique plots were sampled, while the total number of samples taken across the 12 plots was 188.

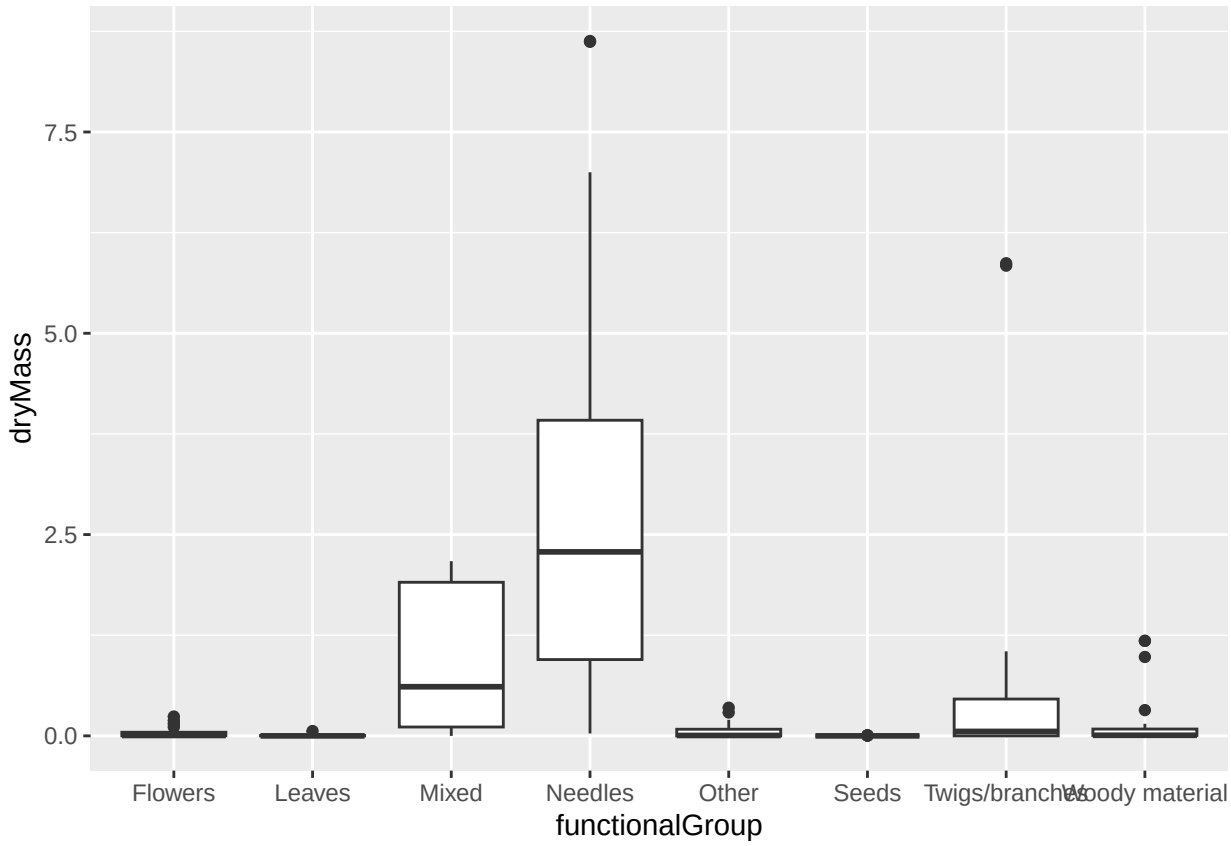
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
# Create bar graph for function group counts in the data set
ggplot(Litter,
       aes( x = functionalGroup)) + geom_bar() + theme(axis.text.x = element_text(angle = 90,
      vjust = 0.5, hjust=1))
```

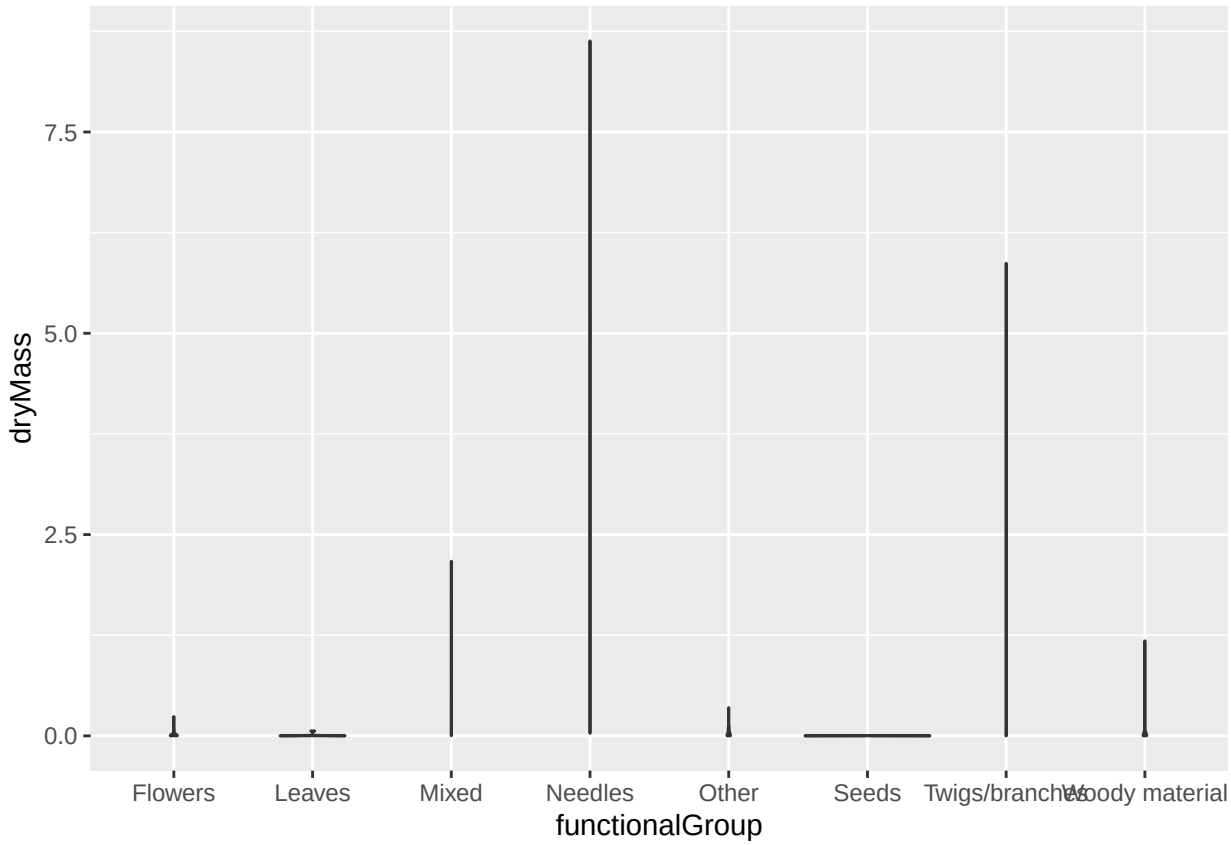


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#Create boxplot
ggplot(Litter, aes( x = functionalGroup, y = dryMass)) + geom_boxplot()
```




```
#Create violin plot
ggplot(Litter, aes( x = functionalGroup, y = dryMass)) + geom_violin()
```



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

> Answer: The box plot is a better method of visualizing this data as it provides more information on the summary, statistics, and outliers. The distribution of the data in this dataset is not suitable for a violin plot due to the outliers affecting the range and distorting the graph, resulting in hard to interpret data. This is likely due to the fact that each category has a relatively small sample size.

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

> Answer: The litter types with highest biomass are 'Needles' and 'Mixed', respectively. The boxplot shows that the mean values of both these categories are the only ones significantly above the zero mark.