Assignment 3: Data Exploration

Sarah Kear

Spring 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 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.

```
getwd()
```

```
## [1] "/Users/sarah/Documents/872_EDA/EDA_Spring2024"
```

```
library(tidyverse)
library (lubridate)

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: There may be interest in the ecotoxicology of neonicotinoid although they are used in agriculture to as a type of insecticide they do have the ability to move to other insects that are not a threat to crops causing them to die. As noted in a recent PNAS article, neonicotinoid poses a broader risk to biodiversity and food webs (https://www.pnas.org/doi/10.1073/pnas.2017221117). It's pertinant that we understand how neonicotinoid, a toxic chemical, impacts our greater ecosystem.

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: Litter and woody debris are an important part of forest and streams ecosystem because it is a food source and habitat for wildlife that live along the streams as well as the accumulation of debris creates carbon sequestration making the ecosystem a carbon sink. There may be interest in studying because it can show the ecosystem, biodiversity, and the overall health of the ecosystem.

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. Litter and fine woody debris are collected using elevated an ground traps. Elevated traps are 80cm off the ground and ground traps are placed on the forest floor. 2. Trap placement is either random or targeted, and is dependent on the type of vegetation. Areas with more than 50% aerial cover of vegetation that is greater than 2m, trap placement is randomized. 3. Ground traps are sampled once per year. In deciduous forest sites, elevated traps are sampled once every 2 weeks during senescence while elevated traps in evergreen sites are sampeld once every 1 to 2 months.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
dimensions <- dim(Neonics)
print(dimensions)
```

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

```r
# the dimensions of the Neonic dataset is 4623 observations and 30 variable
```

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?

```r
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
```

Answer: The two most common effects studied are population and mortality. Population may be the greatest common effect studied to determine how population is effected by neonicotinoids; similarily, mortalilty may also be a top effect studied since neonicotinoids is a type of insecticide and is known to impact the livelihood of non-targeted species.

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...]

```r
# nested the the summary() inside the sort ()  so that the most common species are listed as the top.
# Used the head() to determine the to psix common species studied.

question_7 <- head(
  sort(
  summary(Neonics$Species.Common.Name),
  decreasing = TRUE), n=6
)
question_7
```

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

Answer: The top six commonly studied species in the dataset include: honey bee, parasitic wasp, buff tailed bublebee, carnoilan honey bee, bumble bee, and other. Bees may be of interest since bees are pollinators and have a high chance of interacting the plants that have neonicotinoids. Interfacing with neonicotinoids is one of the causes of bee population decline. Bees are important to an ecosystem as they help sustain biodiversity by pollinating crops and wild plants.

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?

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

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

Answer: `Conc.1..Author.` class is a factor. It is not numeric because the default for uploading data from a CSV file using `read.csv()` is to turn the data into factors. Futhermore, if I wanted to not have the datset uploade as factor, i would make the `stringAsFactors` equal to `FALSE`.

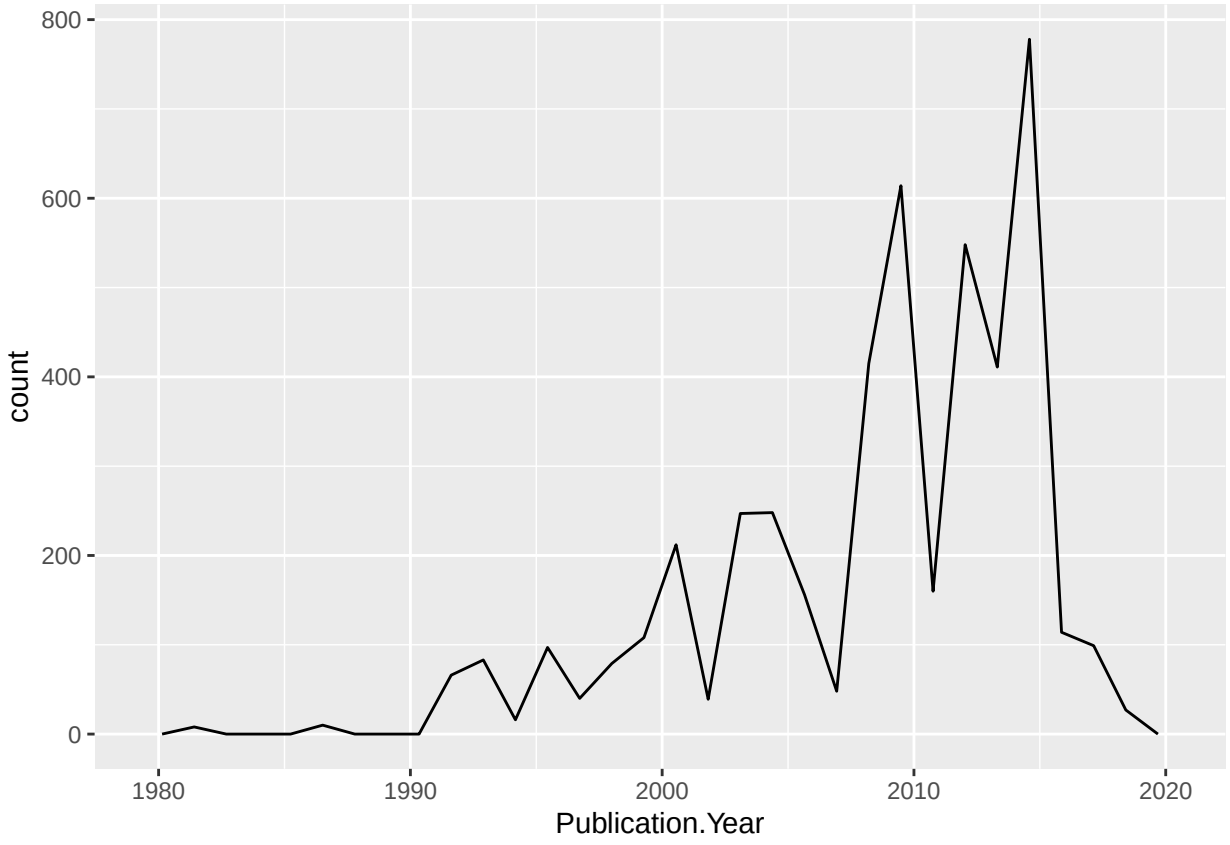
Explore your data graphically (Neonics)

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

```
# used geom_freqpoly() to make graph.

ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year))
```

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

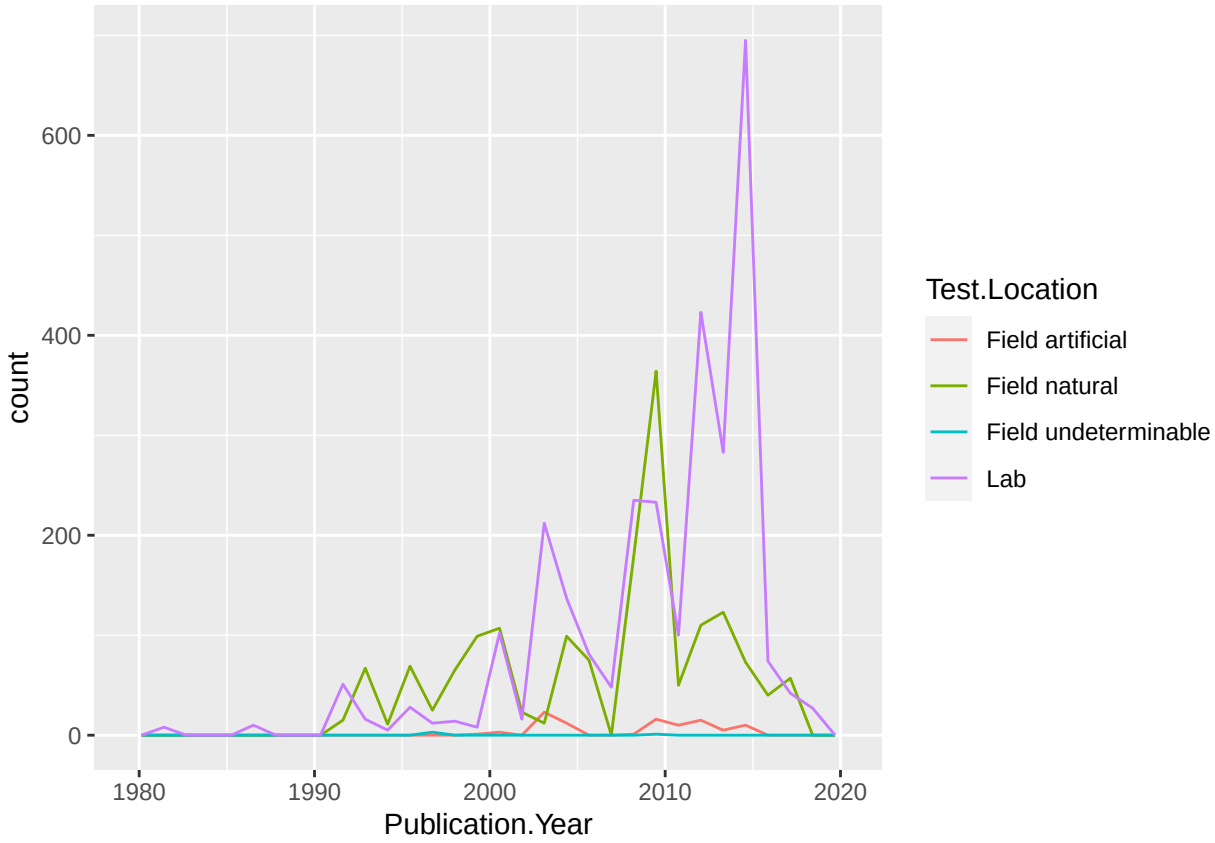


10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
# Determined color inside the aes() to parse out test locations.

ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = 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?

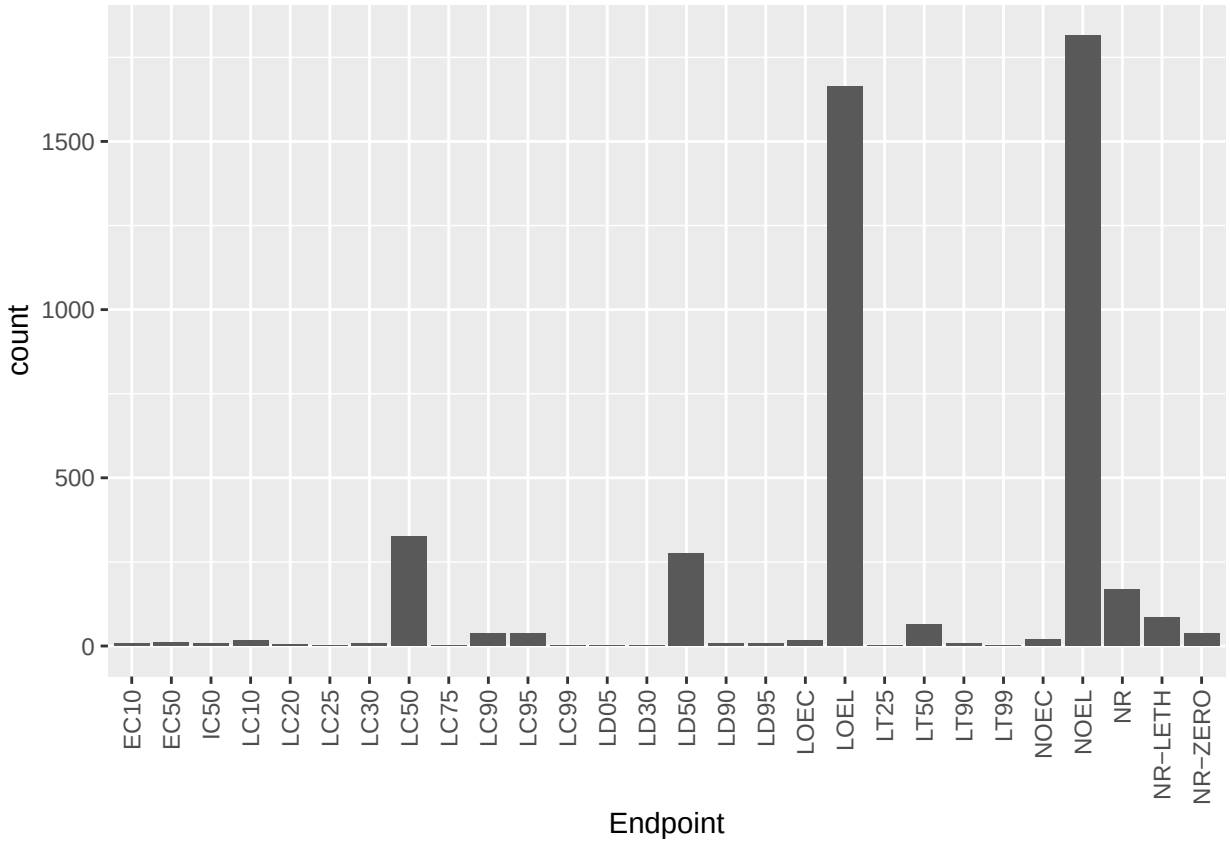
> Answer: The most common test location seems to be the lab, especially between 2000-2005 and 2010-2020. Between 1990-2000 and ~2008 to 2010, the natural field was the most common test location. Overall, the lab test location has continued to increase while the the natural field increased and then has decreased in recent years.

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...]

```
# Used geom_bar() to create bar graph

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 endpoints are LOEL and NOEL. LOEL is from the terrestrial dataset, and stand for Lowest-Observable-Effect-Level, meaning that its effects were different from the control's response. NOEL stands for No-Observable-Effect-Level, meaning the highest concretration's effects are not different from the control.

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 is factor, need to change to Date
class(Litter$collectDate)
```

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

```
#Made a new column and change the format to a date. This needs to be fixed. only showing up as NAs.
Litter$collectDateNew <- as.Date(Litter$collectDate, format = "%Y-%m-%d")

class(Litter$collectDateNew)
```

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

```r
# the two dates litter was sampled was on 8/2/2018 and 8/30/2018
sampledates <- unique(Litter$collectDateNew)
print(sampledates)
```

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

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`?

```r
NiwotRidgeCount <- unique(Litter$plotID)
NiwotRidgeCount
```

```
##  [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
```

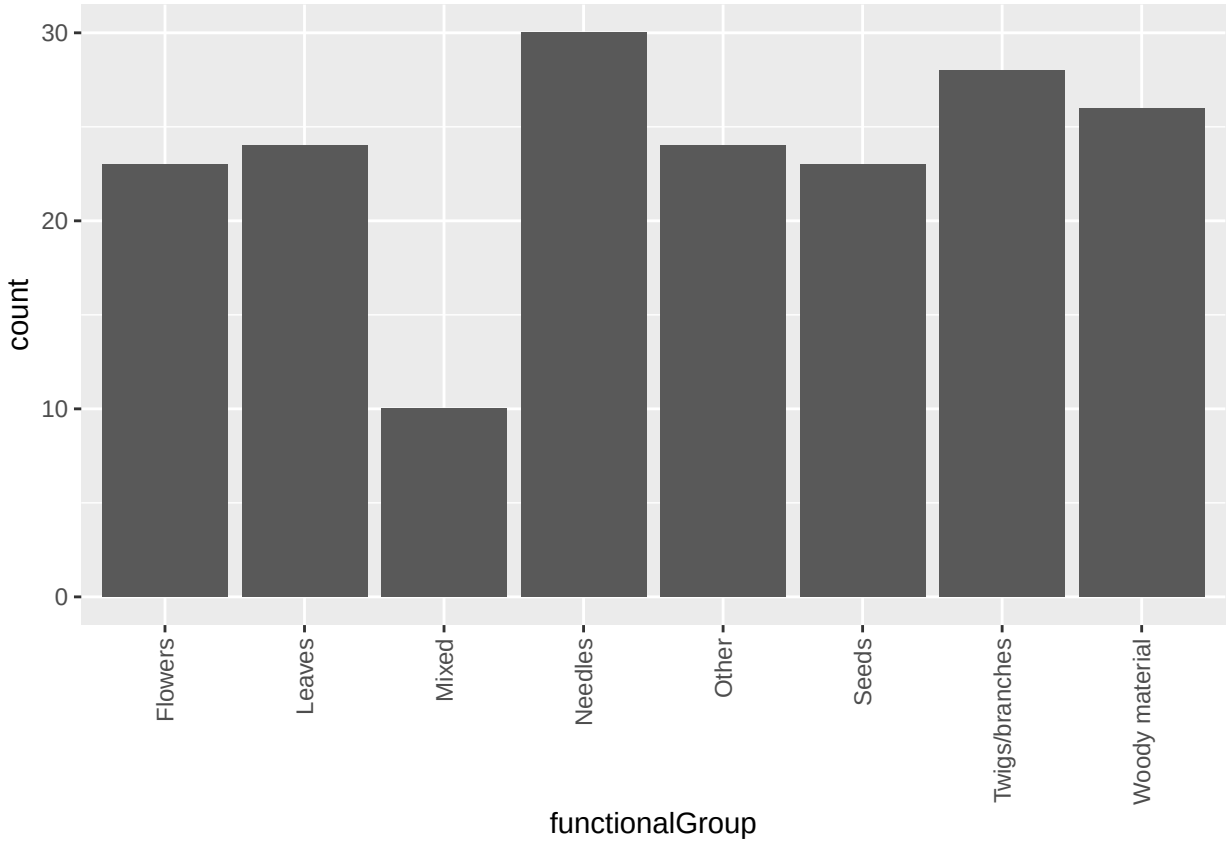
```r
summary(Litter$plotID)
```

```
## 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: 12 plots were sampled at the Niwot Ridge. The information from `unique` differs from the `summary` output because `unique` only outputs the different unique characters, it does not count how many times a unique character has been used. Whereas with `summary` function, it lists each unique character and how many times its been used.

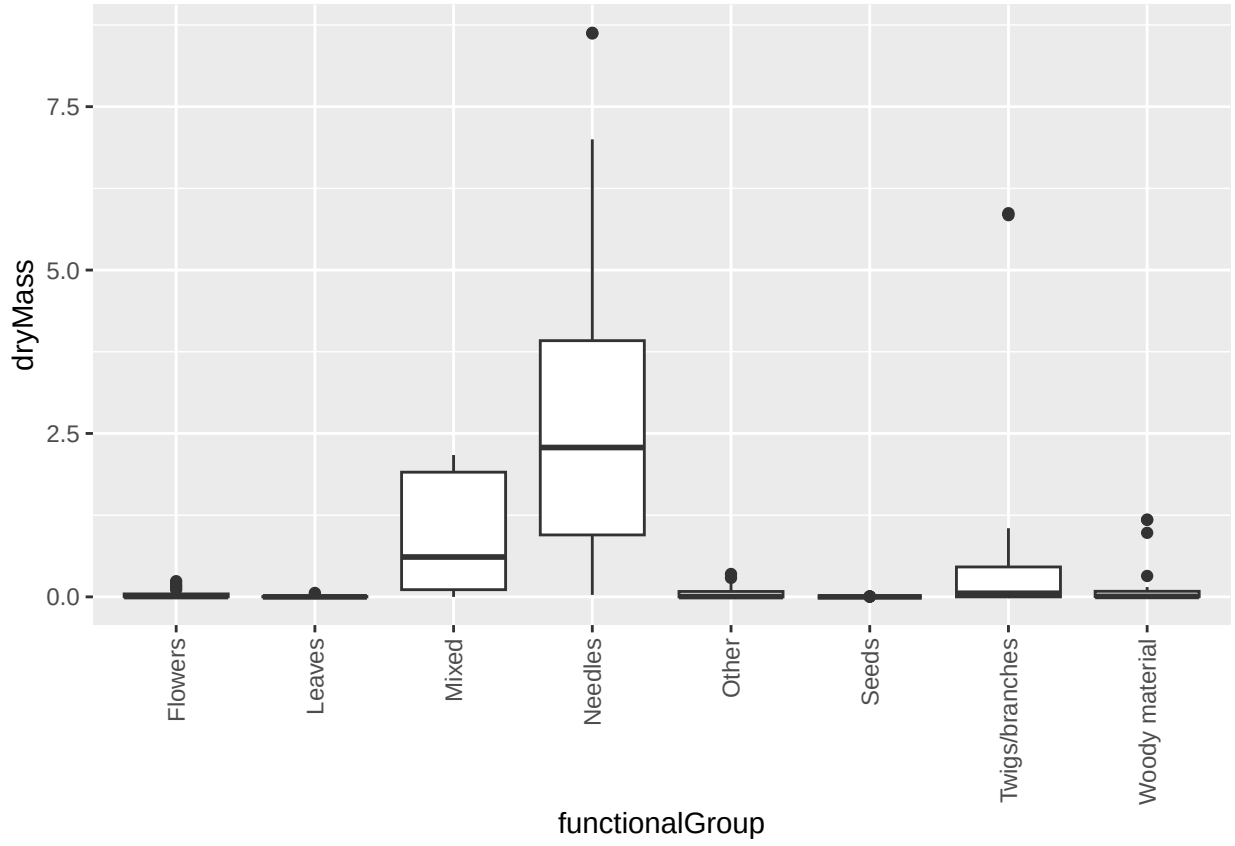
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

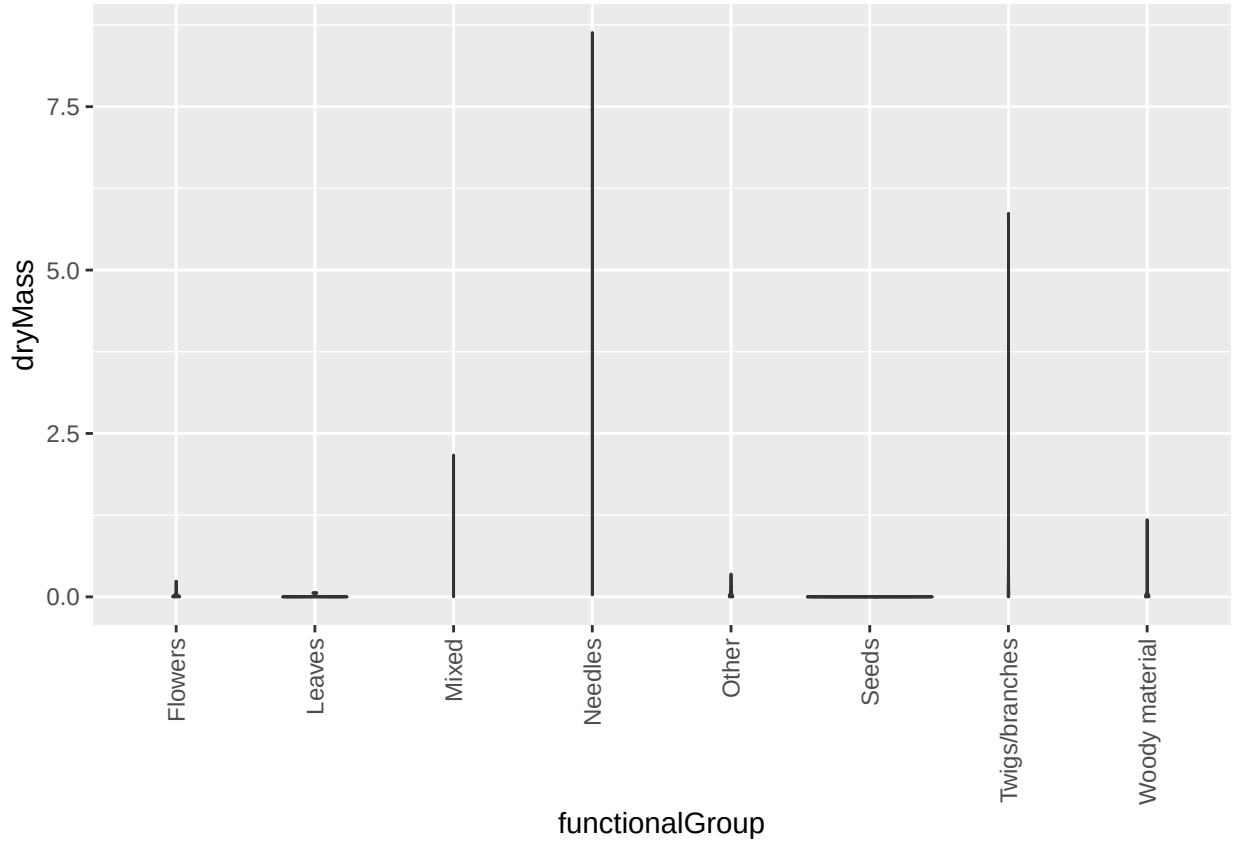


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
# Box Plot
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```




```
# Violin Plot
violinplot <-ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
violinplot
```



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

> Answer: It seems that boxplot is more effective visualization option than the violin plot because there is a lot of variance, such as range and outliers, between different types of litters' biomass. I think if the biomass for each litter were more similar, than the violin plot would have better visualization.

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

> Answer: Based on the box plot, the majority of needles have the highest biomass compared to the other litter types.