Assignment 3: Data Exploration

Kendall Fitzgerald

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.

```
# loading all of my packages
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(here)
```

```
## here() starts at /home/guest/EDE_Fall2024
```

```
getwd() #check working directory
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```
#importing data with relative file path for ecotox neonicotioid dataset and naming it "Neonics"
Neonics <- read.csv(
  file = here("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"),
  stringsAsFactors = TRUE)

#importing data with relative file path for Niwot Ridge NIWON dataset and naming it "Litter"
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: Because Neonicotinoids are widely used as an insecticide, they negatively affect a wide range of insects that could potentially be harmful or damaging to crops. While they target insects that would eat crops, they can also be harmful to insects that are extremely important in pollinating crops like bees. It's important to know how toxic they are to all species of insects in order to determine if they could have potentially adverse consequences on species like bees that are crucial to the survival of many crops.

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 to the ground can be critical in determining many factors about habitats. Litter can have extremely harmful effects on any wildlife that considers Niwot Ridge its home while woody debris plays an important role in nutrient cycling, can be a potential energy source or home for different species, and could influence water flow and sediment transport, as well.

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. Data is only collected at terrestrial NEON sites that contain woody vegetation >2 meters tall. 2. Ground traps are sampled once per year while sampling for elevated traps varies based on vegetation type. 3. Plot edges must be separated by a distance of greater than 150% of one edge of the plot.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
nrow(Neonics) #number of rows in "Neonics" dataset
```

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

```
length(Neonics) #number of columns or "Length" in "Neonics" dataset
```

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

```
dim(Neonics) #dimensions (number of columns and rows) in "Neonics" dataset
```

```
## [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...]

```
#using 'sort' function to put the 'summary' of 'Effects' column into ascending order to determine the l
sort(summary(Neonics$Effect), decreasing = FALSE)
```

```
##       Hormone(s)        Histology       Physiology          Cell(s)
##                1                5                7                9
##     Biochemistry     Accumulation     Intoxication    Immunological
##               11               12               12               16
##       Morphology           Growth       Enzyme(s)         Genetics
##               22               38               62               82
##        Avoidance      Development     Reproduction Feeding behavior
##              102              136              197              255
##         Behavior        Mortality       Population
##              360             1493             1803
```

   Answer: The top five most common effects that are studied are population, mortality, feeding behavior, and reproduction. These would be of interest because they determine the most important indicators of a species' overall health and whether or not they are being negatively affected by the harmful insecticides being used.

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...]

```
#using 'summary' function and setting a maximum number of factors as 7 to show the top 6 species being
summary(Neonics$Species.Common.Name, maxsum = 7)
```

```
##             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 common species that are being studied are the Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. These species are all related to each other as types of bees and play valuable roles in pollinating plants. This pollination factor is most likely the reason why they are being studied more than other insects because of the potential devastation that all crops and plants would feel if these bees died off as a result of the harmful insecticides.

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 of 'Conc.1..Author.' column in dataset
class(Neonics$Conc.1..Author.)
```

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

```
#Viewing data frame, commented out so knitting can happen
# view(Neonics)
```

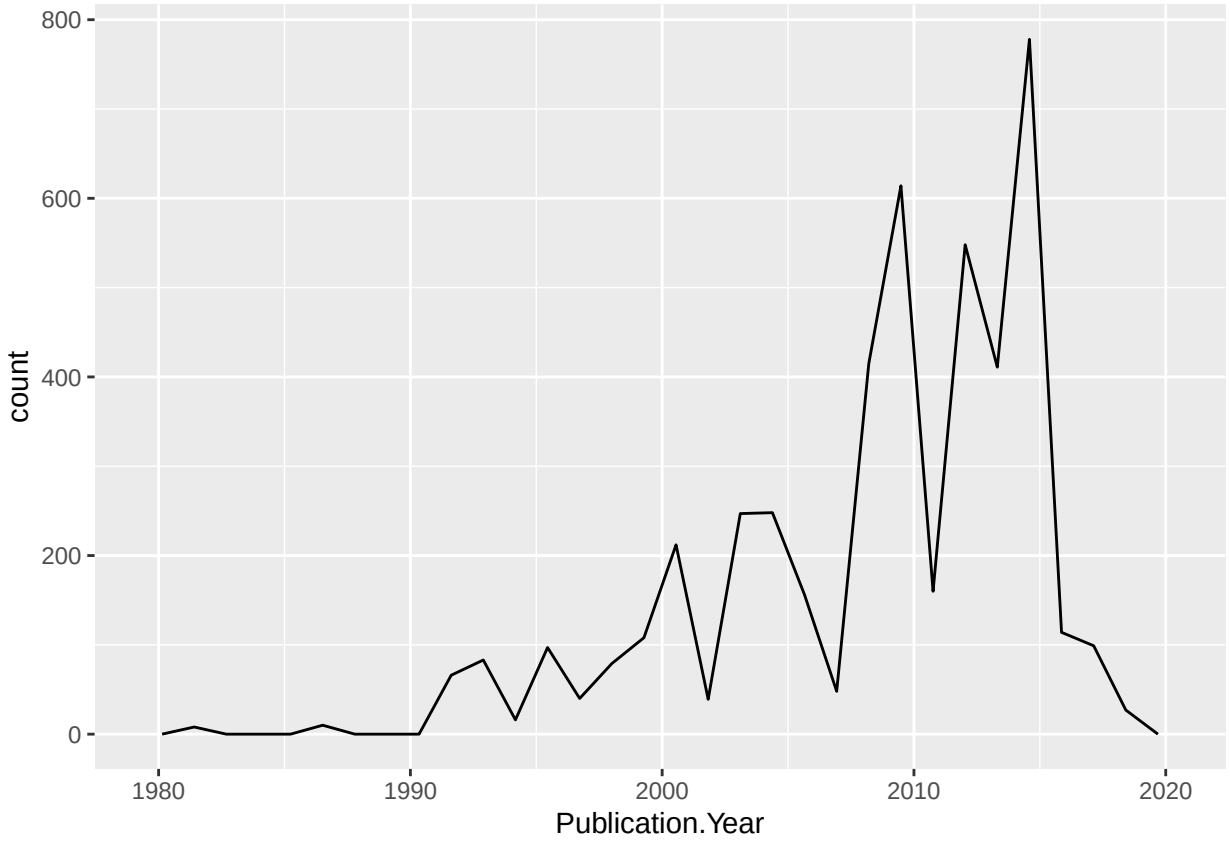
Answer: The class of this column is a "factor". It is not a numeric concentration because there are slashes ('/') in some of the data rows, which would not allow for a numberic classification.

Explore your data graphically (Neonics)

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

```
# Called in ggplot to tell R to use Neonics dataset and set x-axis as "Publication.Year" and then used
ggplot(data = Neonics, aes(x = Publication.Year)) +
  geom_freqpoly()
```

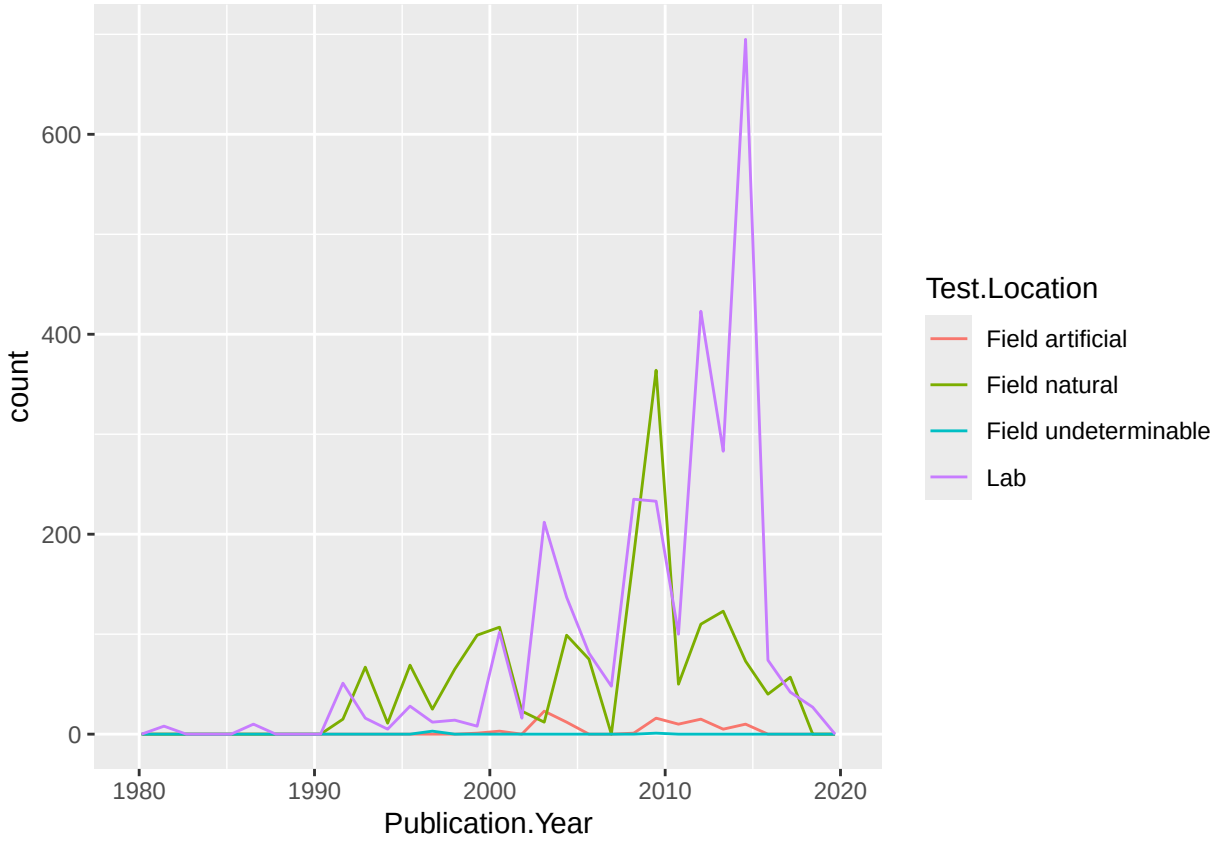
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
# I am using aes to apply a color to each Test.Location and plot that information over the course of pu
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location))
```

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



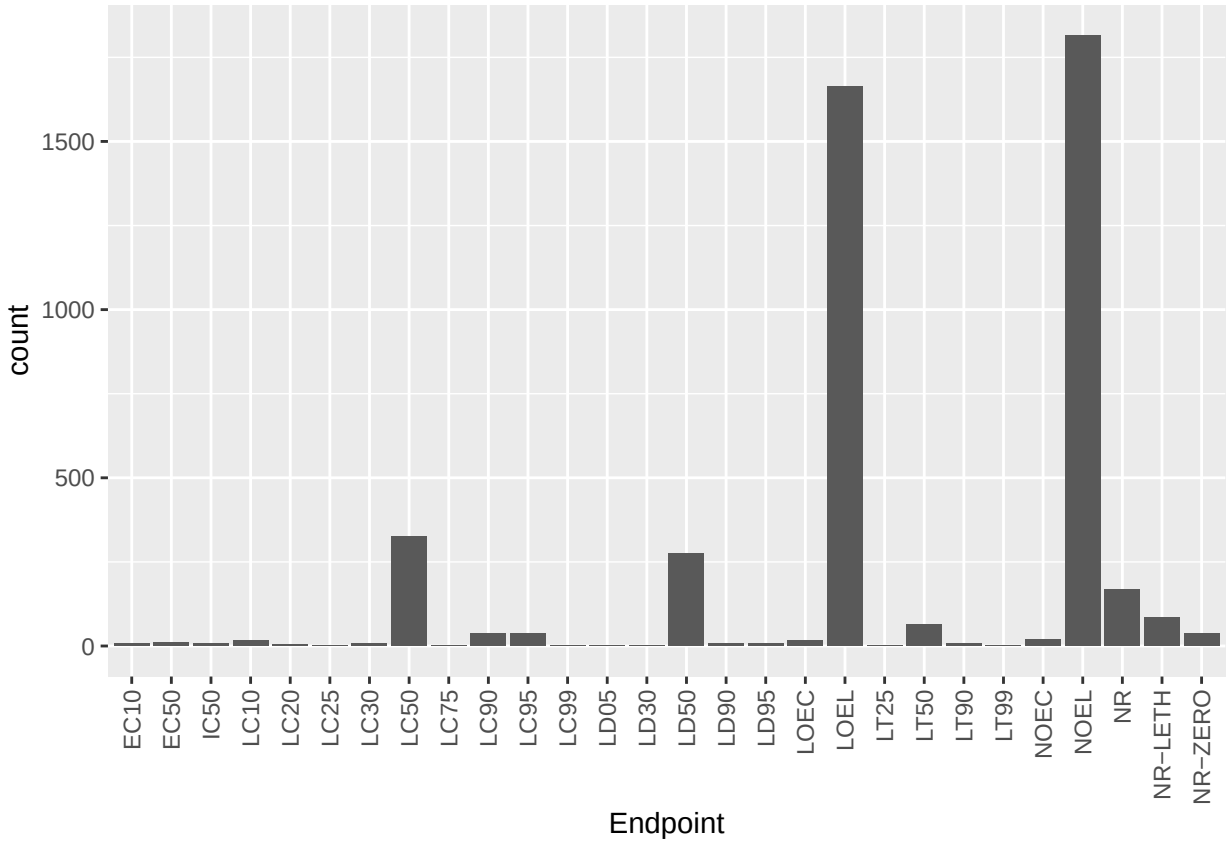
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer:The most common test locations in later years are by far the lab, which experineced a dramatic uptick between 2000 and 2010 and experiencing a sharp decline after 2015.From 1980 to 1990, "field undeterminable" and "lab" were about the same with very little data. "Field natural" experienced a sharp uptick from 2005 to around 2008 and then experienced a drop-off after 2010. "Field artificial" was the least common test location throughout the course of the publication 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…]

```
# I have created a bar graph of all "Endpoint" counts in the Neonics dataset. I have also implemented t
ggplot(data = 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 NOEL and LOEL. NOEL is defined as "No-observable-effect-level" and LOEL is defined as "Lowest-observable-effect-level".

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 of collectDate (it's a 'factor)
class(Litter$collectDate)
```

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

```
#Changing class of collectDate from a 'factor' to a 'date'
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")

#Confirming class of collectDate has been changed to 'date'
class(Litter$collectDate)
```

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

```
#Unique function used to determine which dates litter was sampled in August 2018 (only two dates in the
unique(Litter$collectDate)
```

```
## [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`?

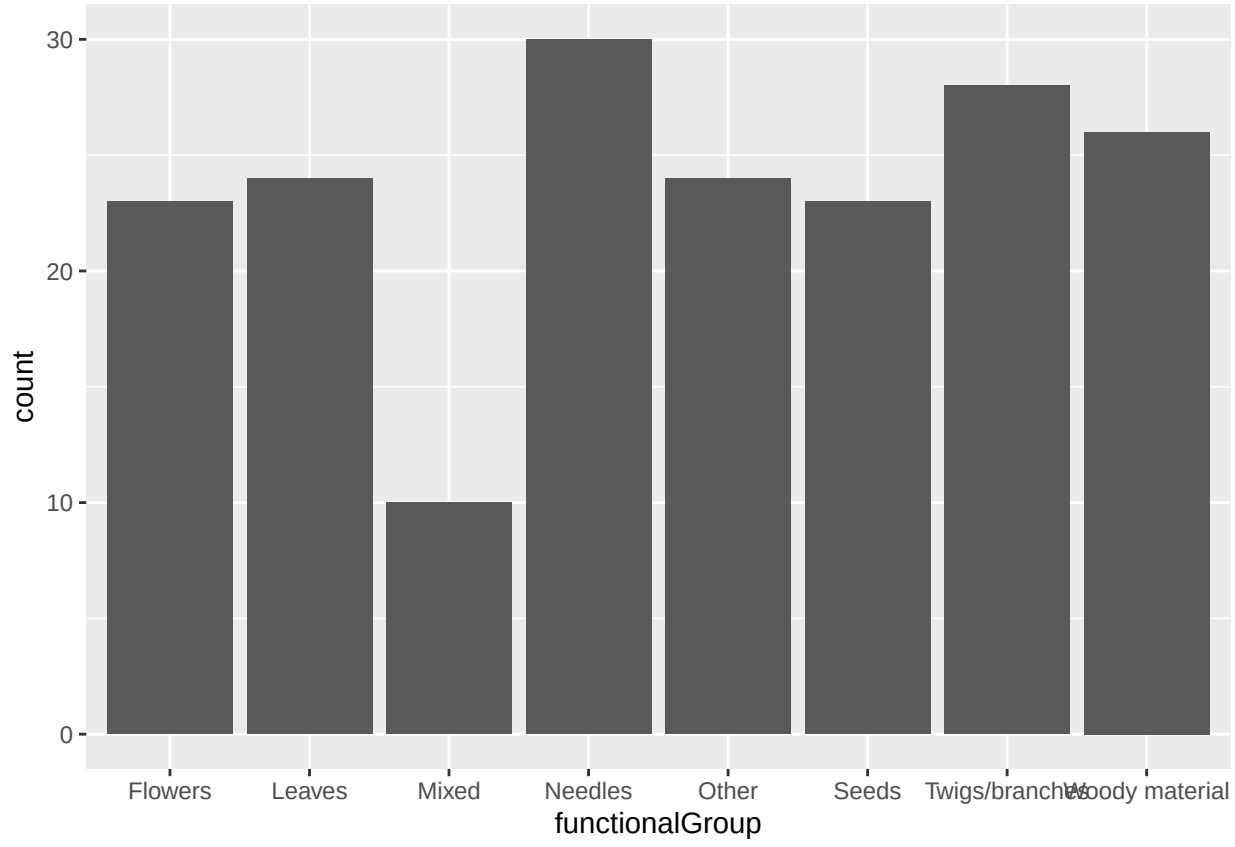
```
#Using 'Unique' function to determine number of plots sampled at Niwot Ridge
unique(Litter$plotID)
```

```
##  [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
```

Answer: Information from the 'Unique' function comes out more compact than information from the 'Summary' Function. The 'Summary' function forces you to count out how many unique values there are yourself while the 'Unique' function says the actual number of different values so that you don't have to count them.

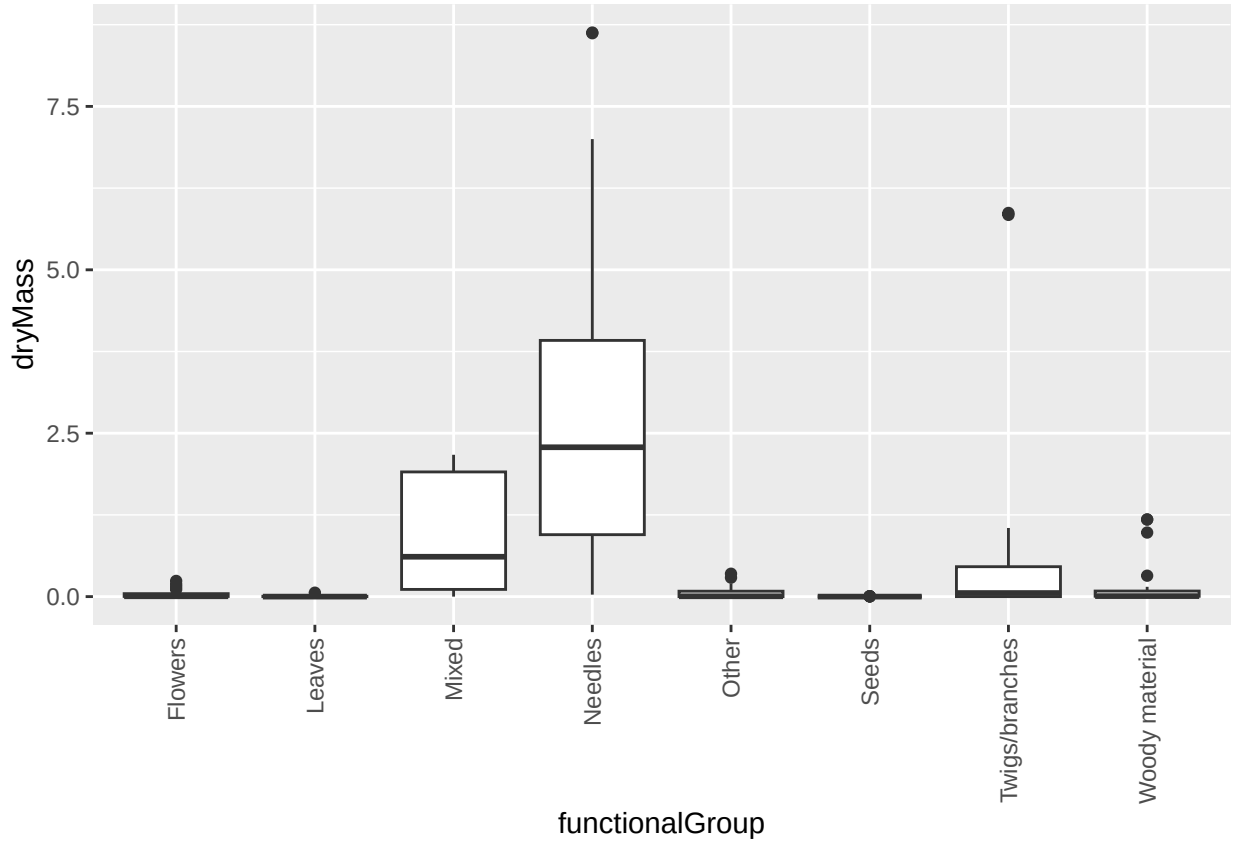
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#Using ggplot to create a bar graph showing the distribution of litter ('functionalGroup' column) acros
ggplot(data = Litter, aes(x = functionalGroup)) +
  geom_bar()
```

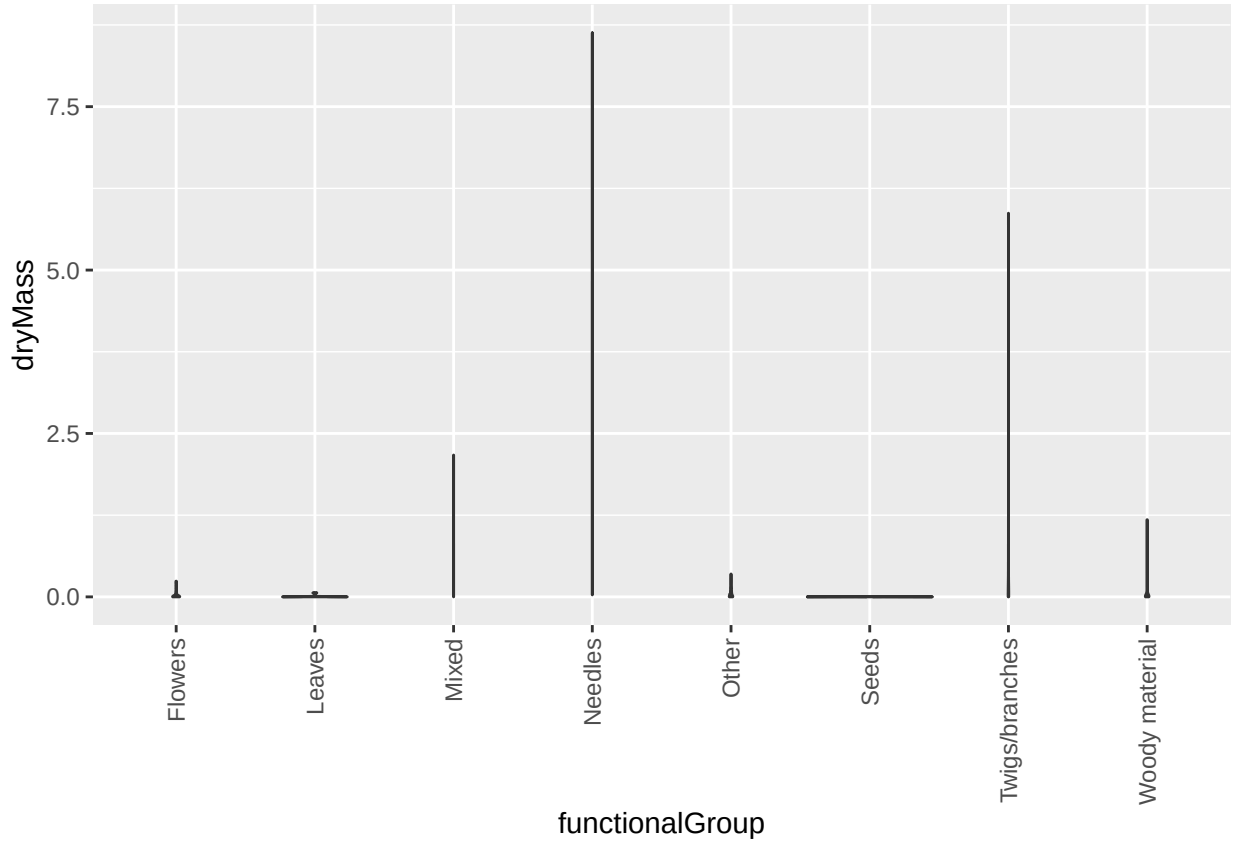


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#Creating a boxplot of the Dry Mass of different kinds of Litter (functionalGroup) and added theme code
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```




```
#Creating a violin plot of the same information
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass)) +
  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: The boxplot is a more effective visualization than the violin plot because the data does not have good enough distribution to be helpful. Unfortunately, the violin plot ends up looking like a series of vertical and horizontal lines that don't tell enough of a story to be helpful.

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

> Answer: 'Needles', 'Twigs/Branches', and 'Mixed' are the types of litter that tend to have the highest biomass at these sites with 'Needles' being the largest of the three.