Assignment 3: Data Exploration

Emily Yang

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.

```
# load packages
library(tidyverse)
library(here)
library(lubridate)
library(ggplot2)

# check working directory
getwd()
```

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

```r
# load the ECOTOX dataset
Neonics <- read.csv(
  file = here("Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"),
  stringsAsFactors = T
)

#View(Neonics)

# load the NEON dataset
Litter <- read.csv(
  file = here("Data/Raw/NIWO_Litter/NEON_NIWO_Litter_massdata_2018-08_raw.csv"),
  stringsAsFactors = T
)

#View(Litter)
```

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 internet search, Neonicotinoids are the most commonly used insecticides in the world. They effectively kill harmful insects, however, they also gave devestaing effects on beneficial insects, pollinators, and aquatic invertebrates. They can be transmitted in through food chains and are very difficult to degrade (source website: "Understanding Neonicotinoids," Xerces Society of Invertebrate Cnnservation).

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:Forest litter and woody debris are ciucial in the process of organic matter decomposition. The decompostion of the organic layer contribute to the nutrient cycle, providing essential nutrients for plants and other organisms in the ecosystem (source: "Decompostition of Orgainc Matter in Caves," Frontiers). The amount of litterfall and fine woody debris can also be used to estimate annual Abovegrounf Net Primary Productivity (ANPP) and aboveground biomass (metadata).

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. Traps: litter and wood debris were collected using elevated 0.5 m^2 PVS traps and ground traps, respectively. 2. Spatial Sampling: litter and woody debris sampling were conducted at terrestrial NEON sites tower plots. Locations of twoer plots were selected randomly, 40 * 40m and 20*20m plots were created in different locations, and traps were placed within the plots. 3. Temporal Sampling: Ground traps were sampled once a year, and the sampling frequecy for elevated traps varies. Traps in decidious forest were sampled every two weeks during senescence, and traps in evergreen sites were sampled every 1-2 months throughout the year.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) # used dimension function
```

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

```
# there 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? [Tip: The `sort()` command is useful for listing the values in order of magnitude...]

```
sort(summary(Neonics$Effect)) # summazied and sorted the Effect column
```

```
##       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 most common effect studied was Population.The study is interested in studying the effect of neonicotinoid on the insect population. Researchers could study whether neonicotinoid positively or negatively affects the insect population.

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...]

```
# help(summary)
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
```

```
# Used maxsum = 7 because the most studied output is (Other), which does not have a species common name
```

Answer: The six most common species studied are: Honey Bee (677), Parasitic Wasp (285), Buff Tailed Bumblebee (183), Carniolan Honey Bee (152), Bumble Bee (140), Italian Honeybee (113). These species are all pollinators, which are circial for the plant growth.

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...]

```
class(Neonics$Conc.1..Author.) # used class function
```

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

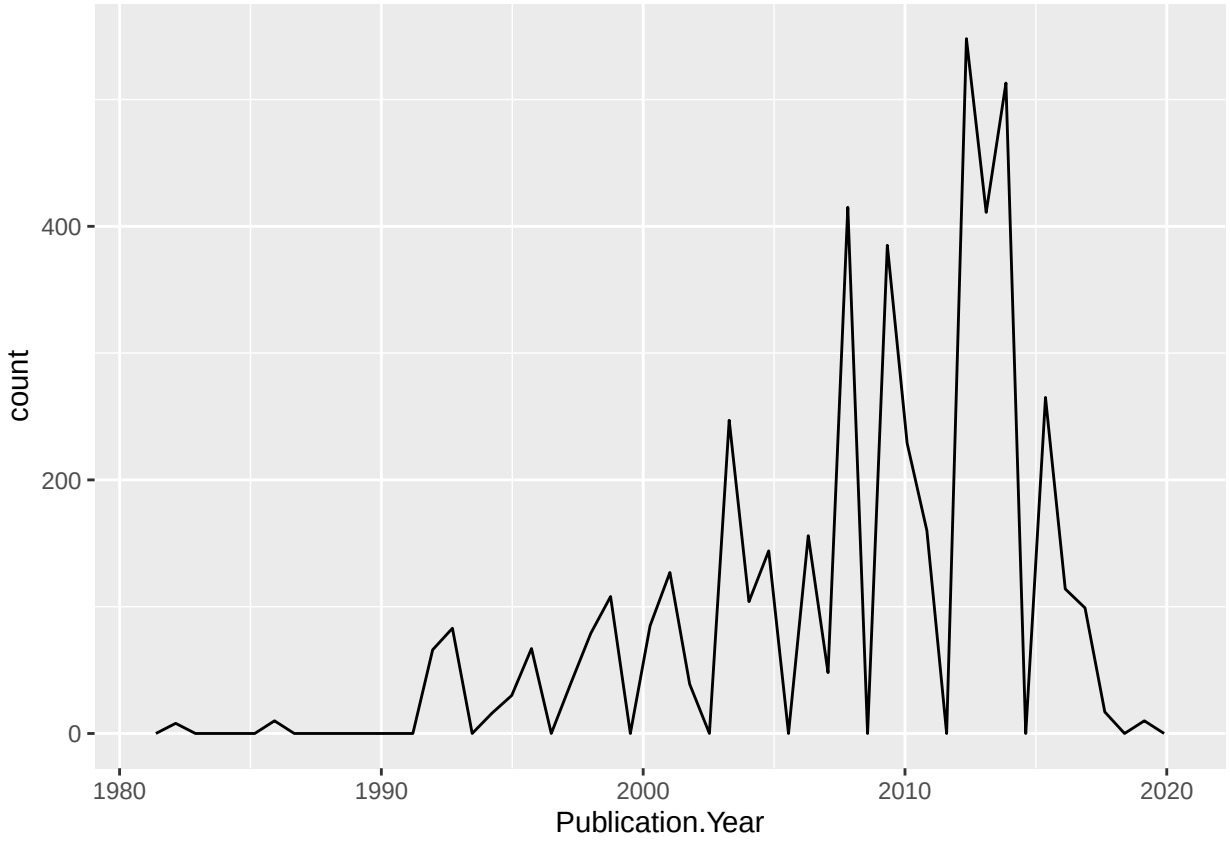
```
#View(Neonics)
```

Answer: The class of 'Conc.1..Author." is factor. By viewing the dataset, I found that some of the variables are approximate (e.g. "~10") and had special characters such as'/' at the end of the number, which caused R to interpret these values as strings rather than numbers. Consequently, when the dataset was imported, this column was automatically converted to factors with the use of the stringsAsFactors parameter.

Explore your data graphically (Neonics)

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

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50)
```

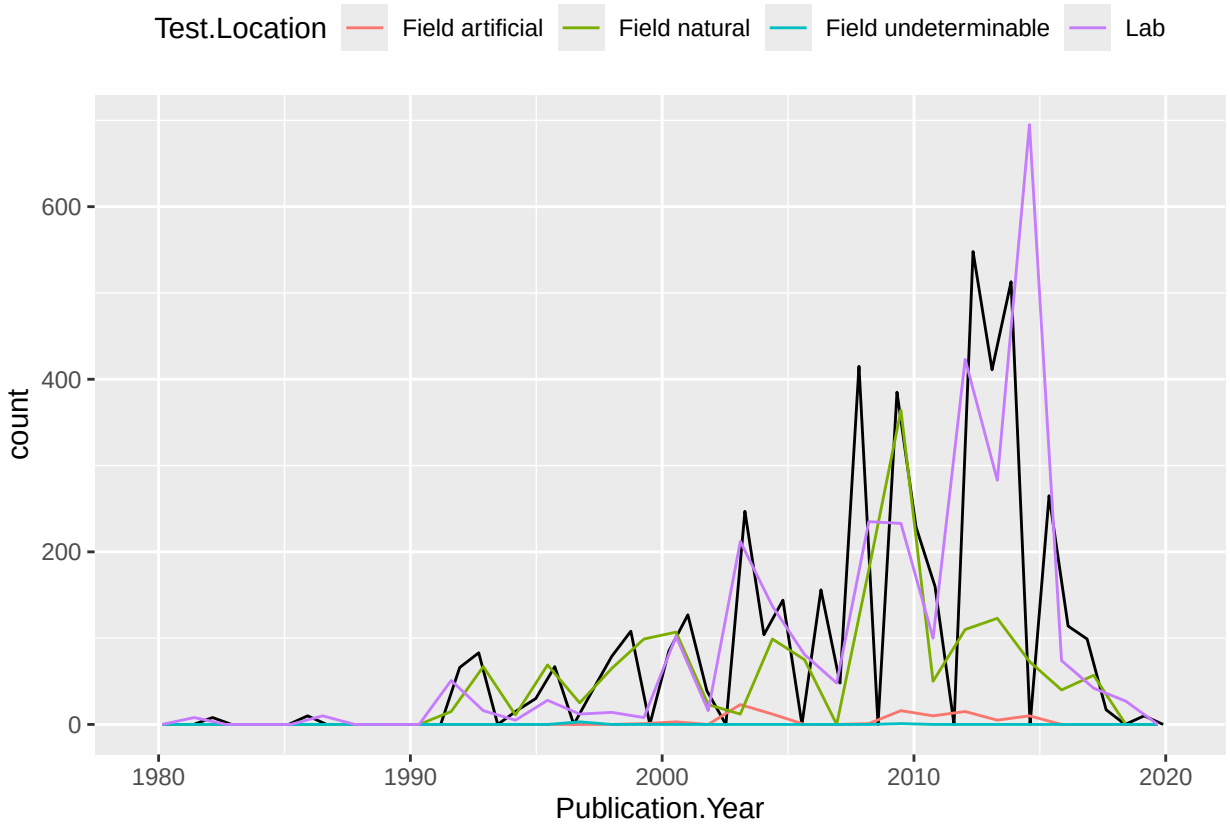


```
#used geom_freqpoly function, and put Publication.Year on the x axis to see the count
```

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

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location)) +
  theme(legend.position = "top")
```

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




```
# added a color aesthetic and top the legend at the top of the graph
```

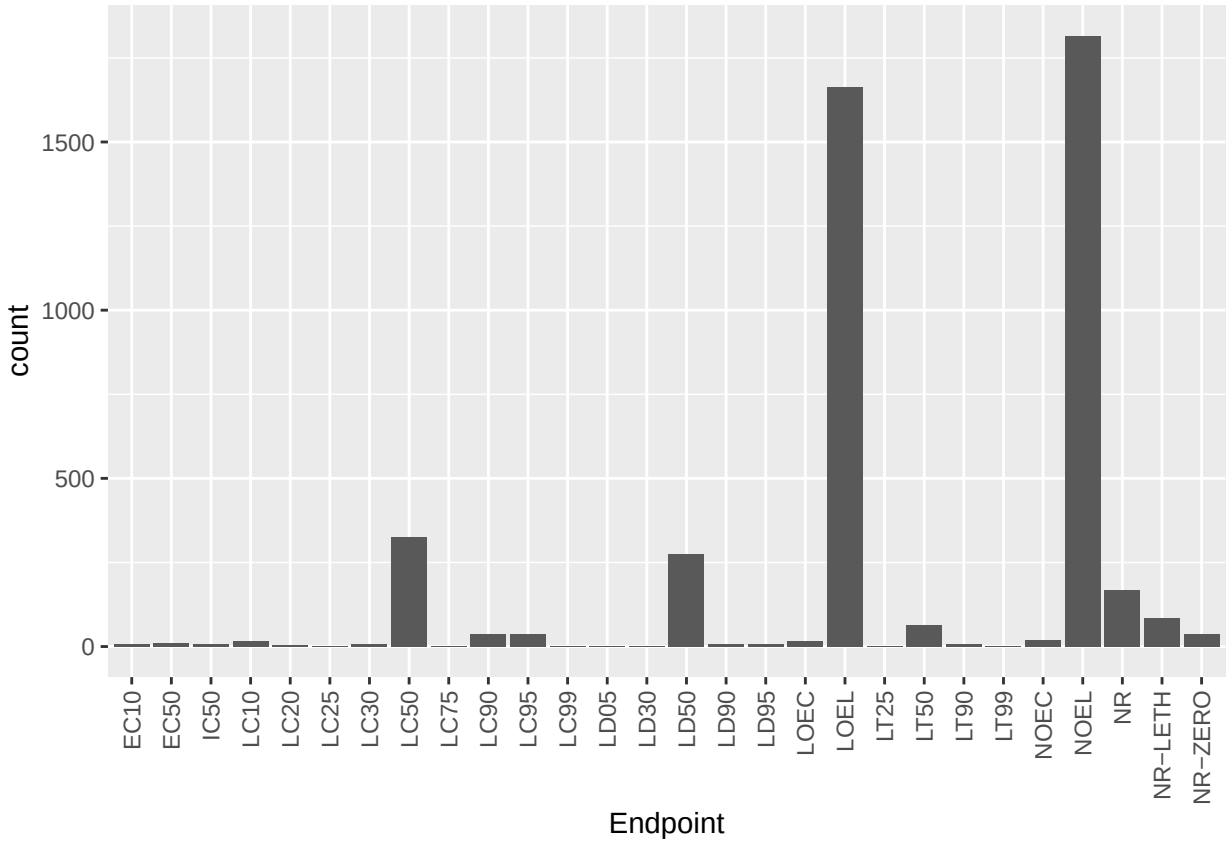
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test lcoations are in labs(purple line). The lab test locations started increasing around 1999, peaked between 2012 to 2016 with over 600 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...]

```
ggplot(data = Neonics, aes(x = Endpoint)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: The most common ones are "LOEL" and "NOEL". According to the metadata, LOEL is identified as Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC). NOEL is No-observable-effect-level: highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test (NOEAL/NOEC). Both are for terrestrial database usage.

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) # shows the class is factor
```

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

```
#head(Litter$collectDate) # view the first few rows

# convert the collectDate from factor to data format
collect_date <- (Litter$collectDate)
collect_date_formatted <- as.Date(collect_date, format = "%Y-%m-%d")
# check the format to ensure it's Date
class(collect_date_formatted)
```

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

```
#help(unique)

# Find the dates litter was sampled in August 2018
unique(collect_date_formatted)
```

```
## [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`?

```
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
```

```
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
```

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

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

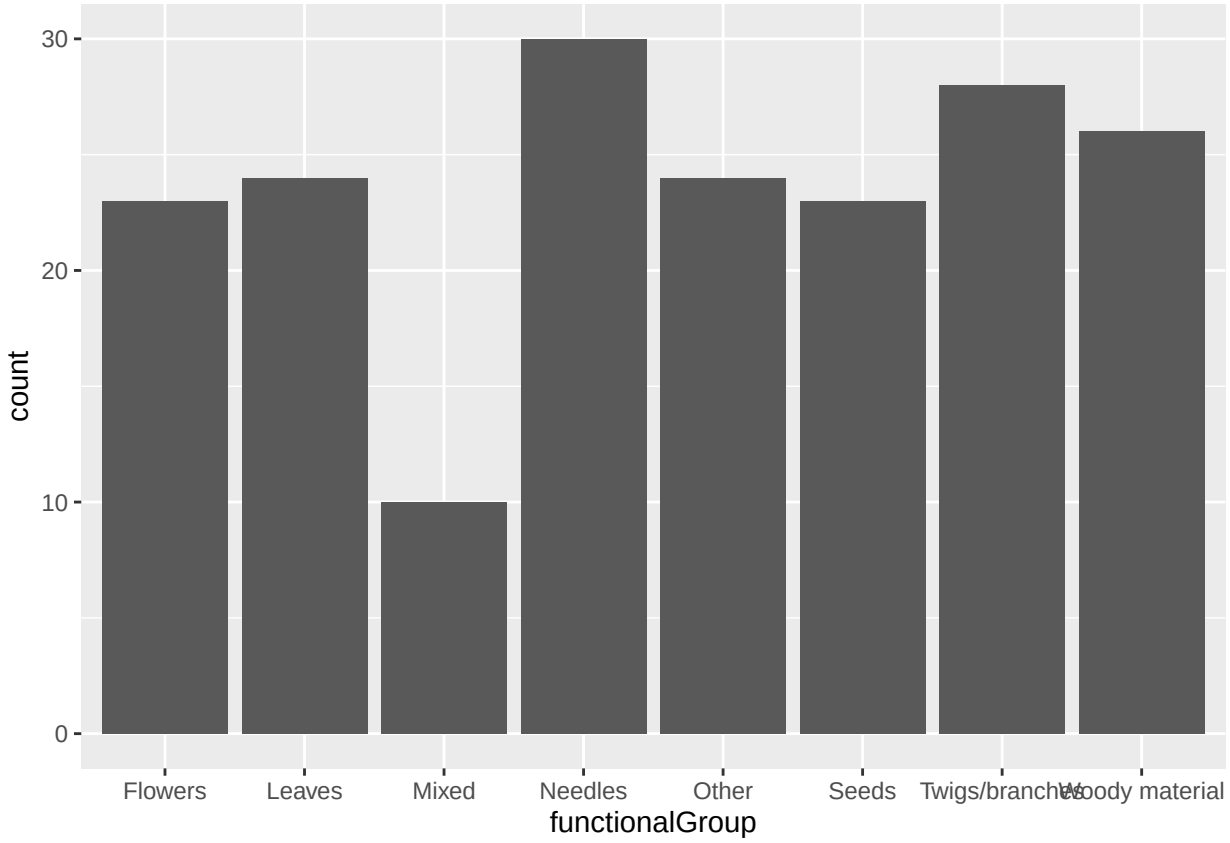
```
length(summary(Litter$plotID))
```

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

Answer: There are 12 different plot IDs sampled at Niwot Ridge. The unique function provides a list of 12 different plot ID without repeating and indicates there are 12 levels. The summary function provides a frequency count for each unique value in the plotID column.

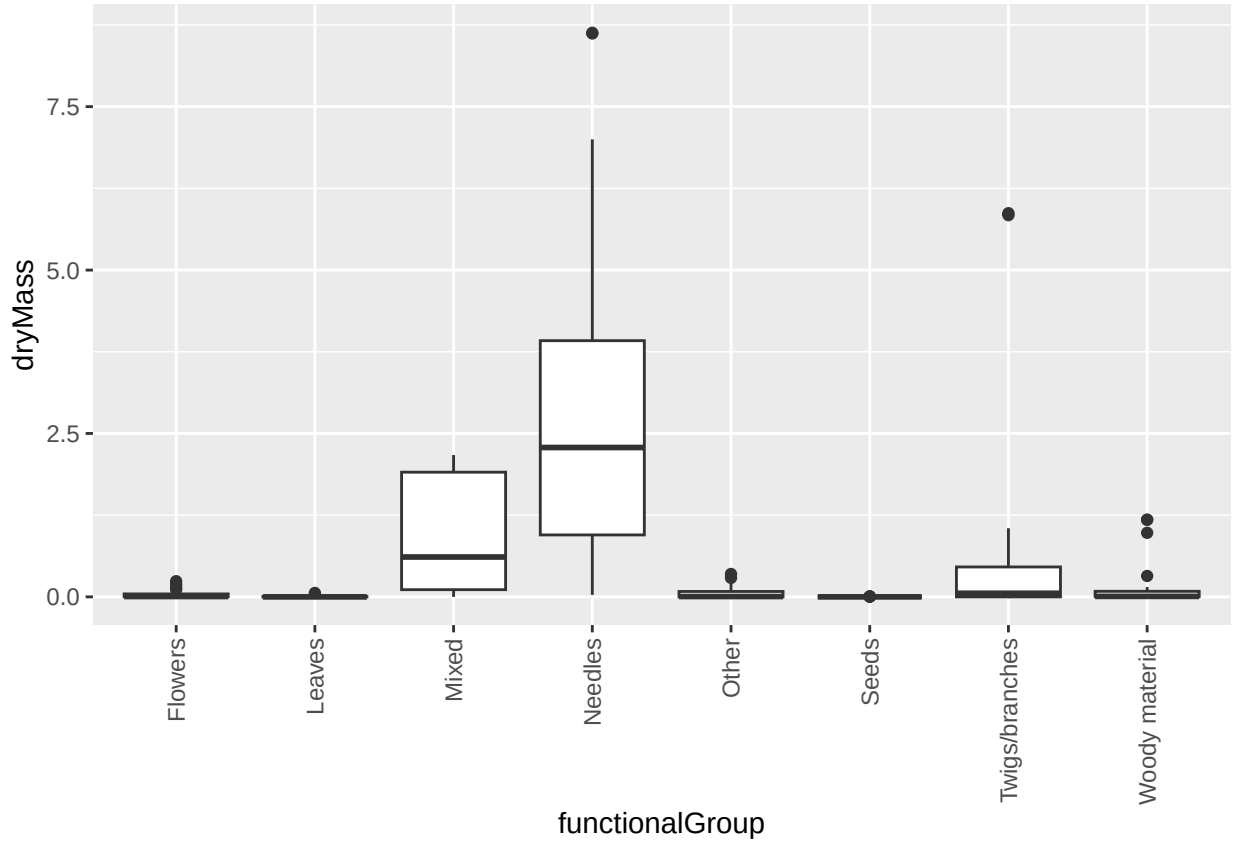
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(data = Litter, aes(x = functionalGroup)) +
  geom_bar()
```

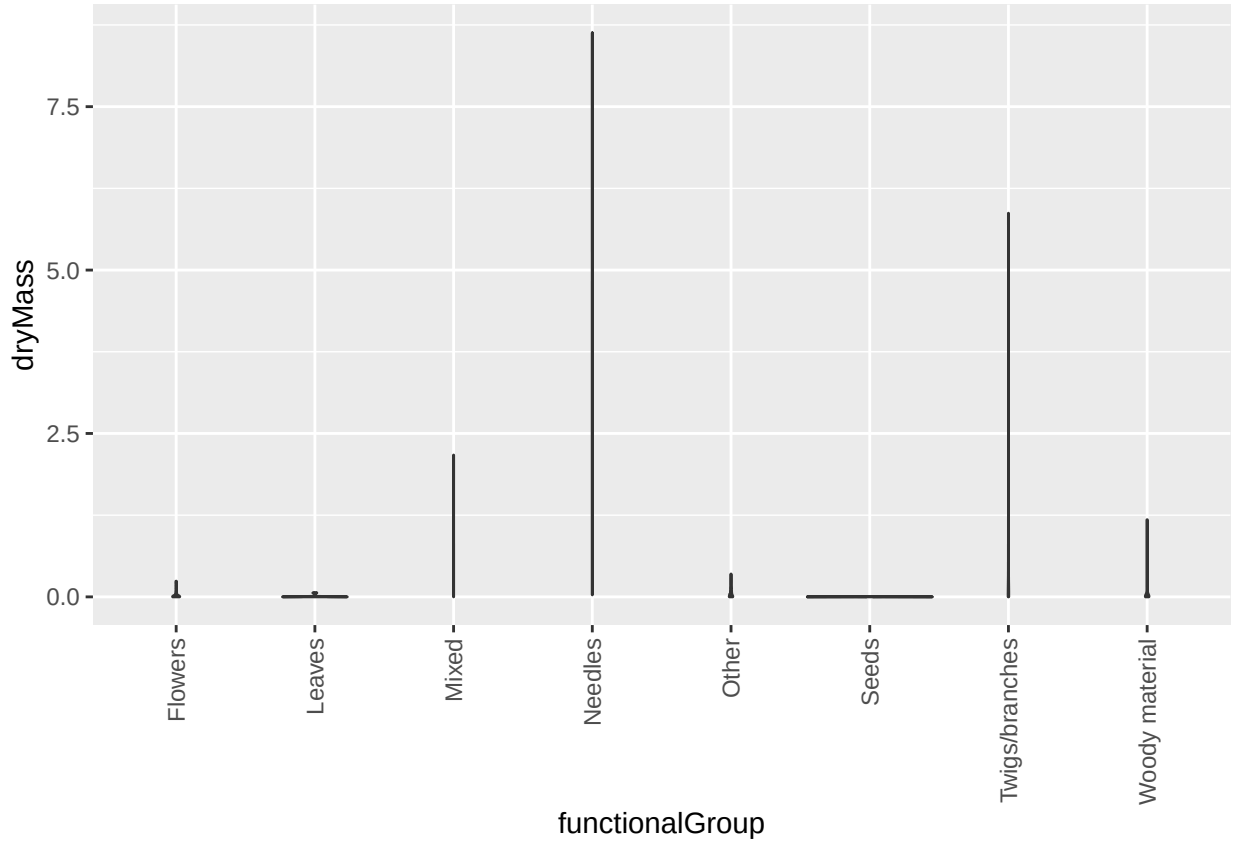


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




```
# Violin plot
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:Boxplot provides more clarity than violin plot in this case, as many of the categories such as “Flowers”,“leaves”, and “seeds” have very small smaple sizes. Boxplot is more effective in showing the distribution and spread, as the density estimation is not effective with violin plots in this case.

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

> Answer: According to the boxplot, “Needles” and “Mixed” litter types have the hightest biomass at these sites.