Assignment 3: Data Exploration

Alex Lopez

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.

```r
#Load packages
library(tidyverse)
library(lubridate)
library(here)

#Check working directory
getwd()
```

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

```
here()
```

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

```
#Import datasets
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: According to the Xerces Society for Invertebrate Conservation, research has shown that neonicotinoids have had significant negative impacts on the environment. Specifically, neonicotinoids "are toxic to pollinators, beneficial insects, and aquatic invertebrates."

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: According to the U.S. Department of Agriculture, litter and woody debris play an important role in carbon budgets and nutrient cycling, are a source of energy for aquatic ecosystems, and provide habitat for terrestrial and aquatic organisms, among other things.

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 sampling is executed at terrestrial NEON sites that contain woody vegetation >2m tall. 2. Trap placement within plots may be either targeted or randomized, depending on the vegetation. 3. Ground traps are sampled once per year.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#Dimensions of the dataset
dim(Neonics)
```

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

```
#The Neonics dataset 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...]

```
#The 'summary' function lists the common effects that are studied and their
#respective counts.
#The 'sort' function lists them in order of magnitude from least to greatest.
sort(summary(Neonics$Effect))
```

```
##       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 effects that are studied are: population and mortality, each with a count greater than 1000. These effects might be specifically of interest, because we already know neonicotinoids are toxic/deadly to certain animals, so it would make sense that we would want to place a greater focus on population and mortality.

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...]

```
#The 'summary' function, with the 'maxsum' argument equal to 7, lists the top 6
#levels of the factor, with the rest being combined into one category called
#'(Other)'.
#So, in our case, the six most commonly studied species in the dataset are
#returned.
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: Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee are the six most commonly studied species in the dataset.

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…]

```
#Check class of 'Conc.1..Author'
class(Neonics$Conc.1..Author.)
```

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

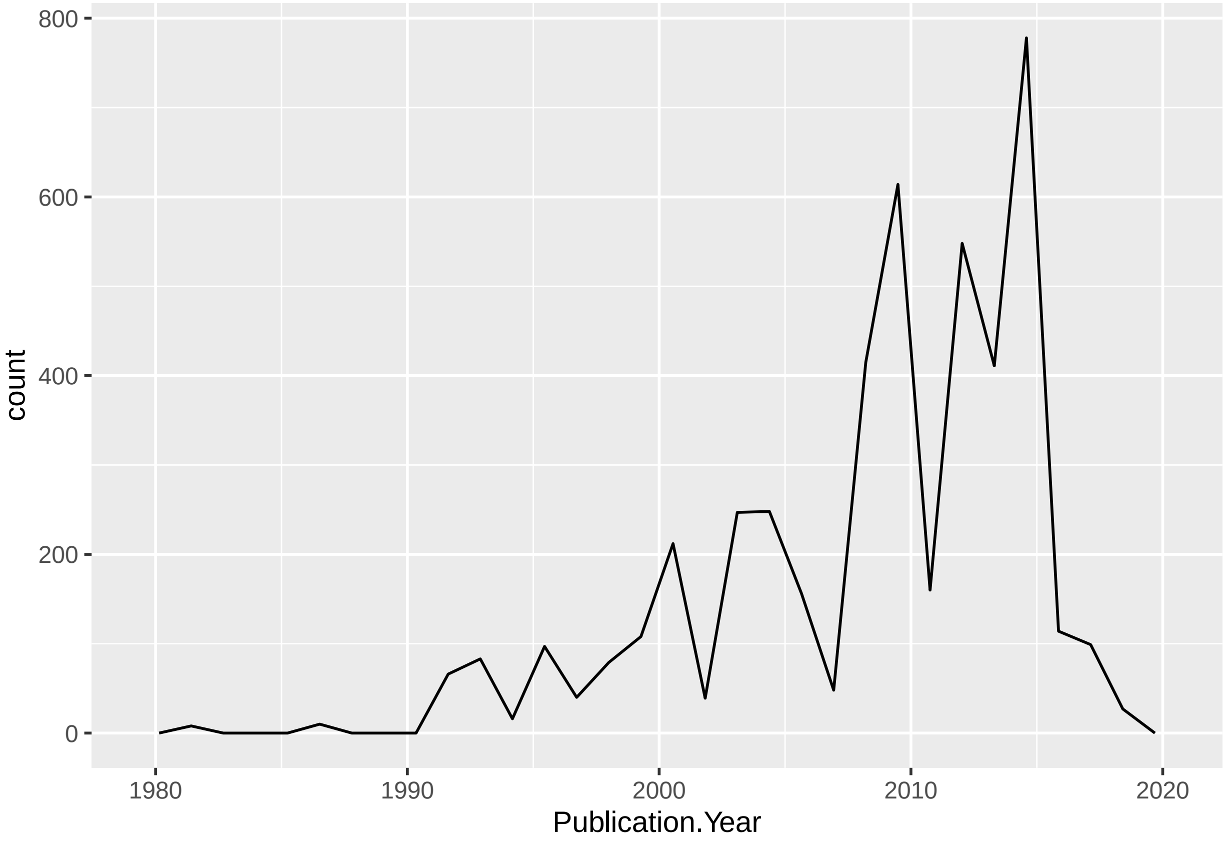
Answer: The class of 'Conc.1..Author' is a factor. It is a factor and not numeric, because the values contain non-numeric characters such as '<'. As a result, when we used 'stringsAsFactors = TRUE' when importing the dataset, R must have interpreted the values of 'Conc.1..Author' as non-numeric and therefore converted them to factors.

Explore your data graphically (Neonics)

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

```
#Use ggplot() and place Publication.Year goes on the x-axis
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year))
```

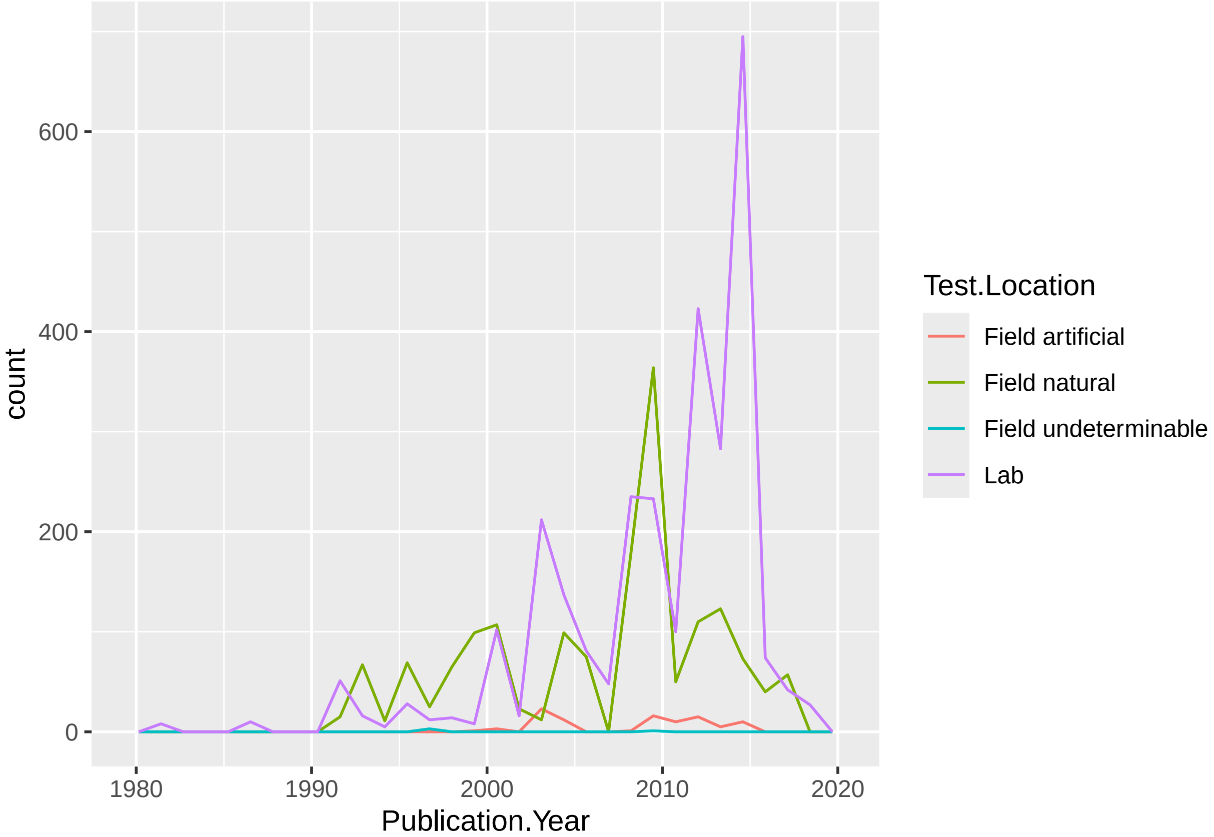
```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



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

```
#Add Test.Location as different colors
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location))
```

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



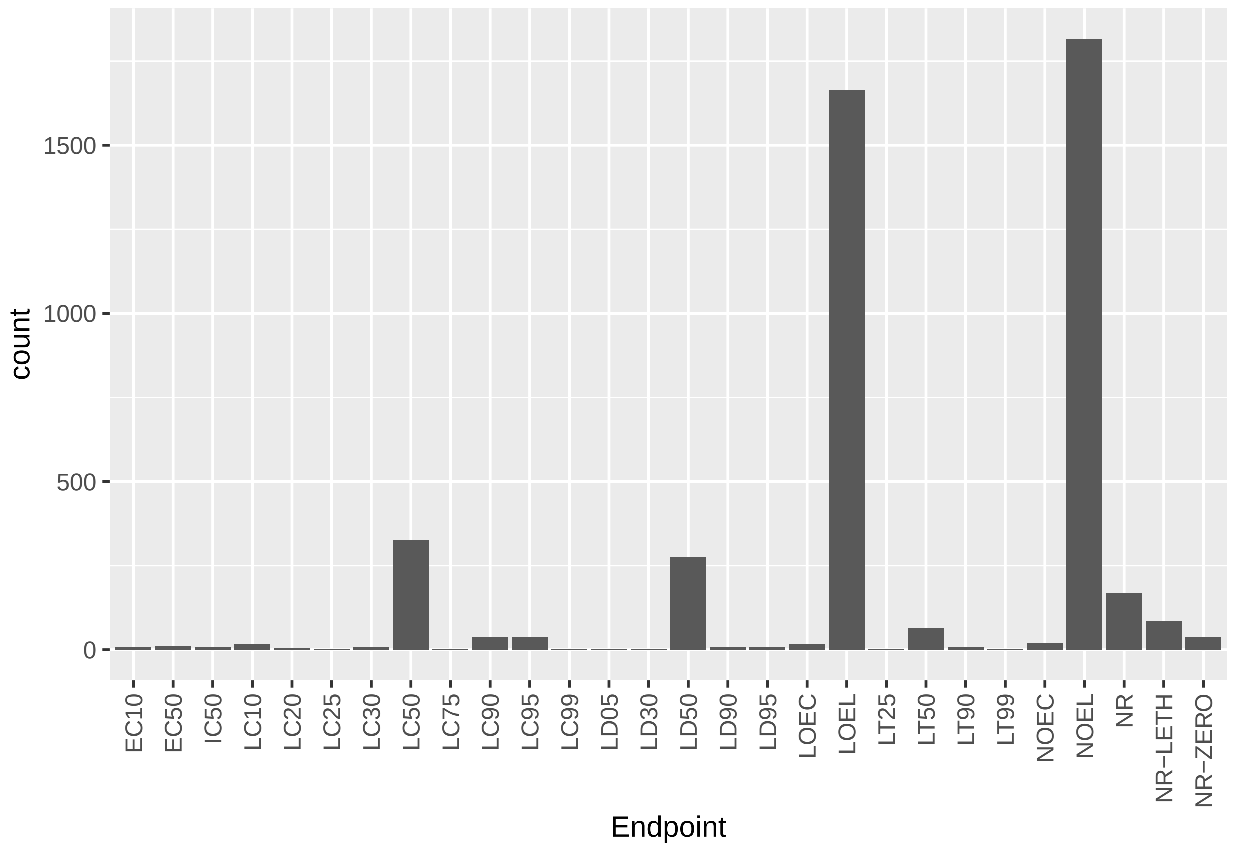
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are ‘Lab’ and ‘Field Natural’, the most common test locations starting from around 1990. Between 2010 and 2020, although they remain the most common test locations, counts of ‘Field Natural’ are decreasing. In the case of the ‘Lab’ test location, counts increase from around 2010 to around 2015, after which they decrease until 2020.

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 two most common endpoints are LOEL and NOEL. LOEL is defined as: Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC), and NOEL is defined as: 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).

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.

```
#Check class of 'collectDate'
class(Litter$collectDate)
```

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

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

#Confirm new class of 'collectDate'
class(Litter$collectDate)
```

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

```
#Use 'unique()' function to determine which dates litter was sampled in Aug 2018
august_dates <- unique(Litter$collectDate[Litter$collectDate >= "2018-08-01" &
                                            Litter$collectDate <= "2018-08-31"])
august_dates
```

```
## [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`?

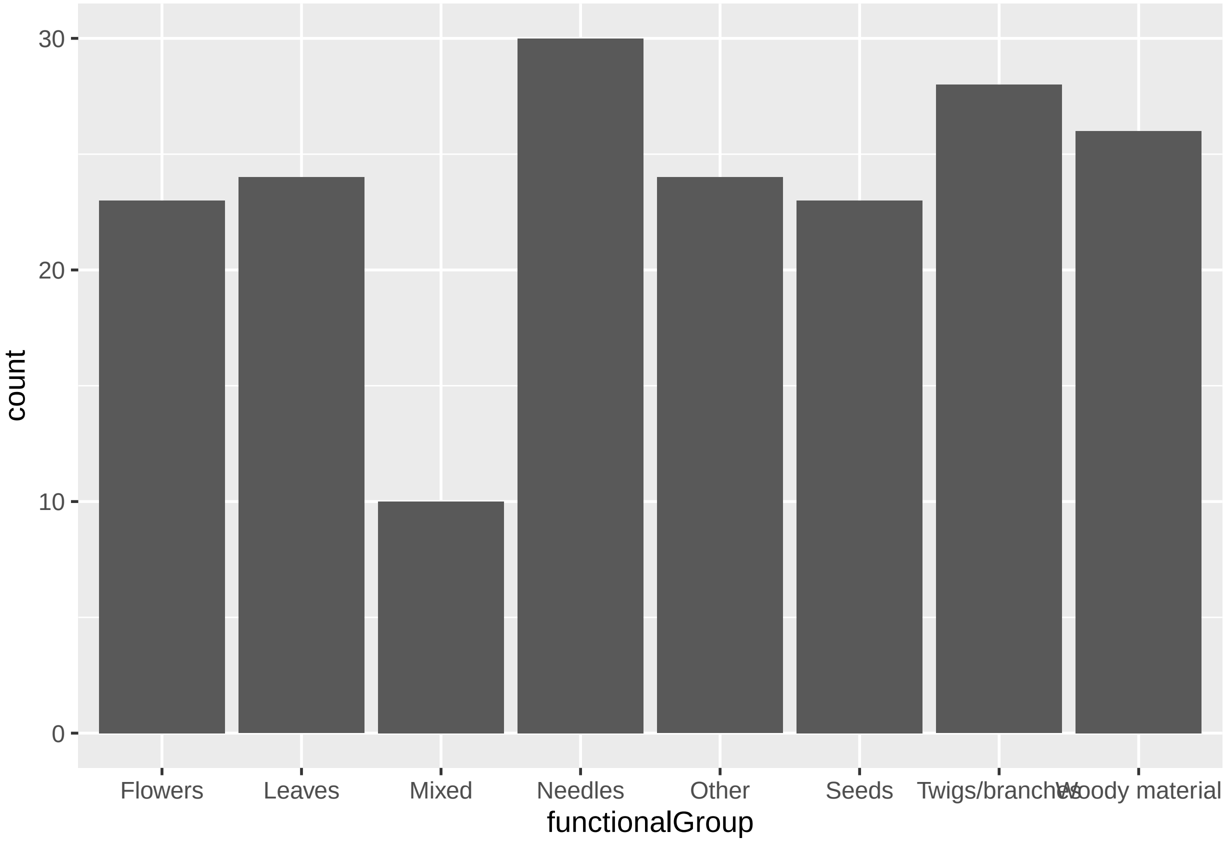
```
#Use unique() function to find distinct plot IDs
unique_plots <- unique(Litter$plotID)

#Use length() to find number of distinct/unique plot IDs
number_unique_plots <- length(unique_plots)

#Print the result
number_unique_plots
```

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

Answer: 12 different plots were sampled at Niwot Ridge. The unique() function returns only the unique/distinct values/levels from a vector or column. It does not provide the frequency or count of each level, which summary() does do.

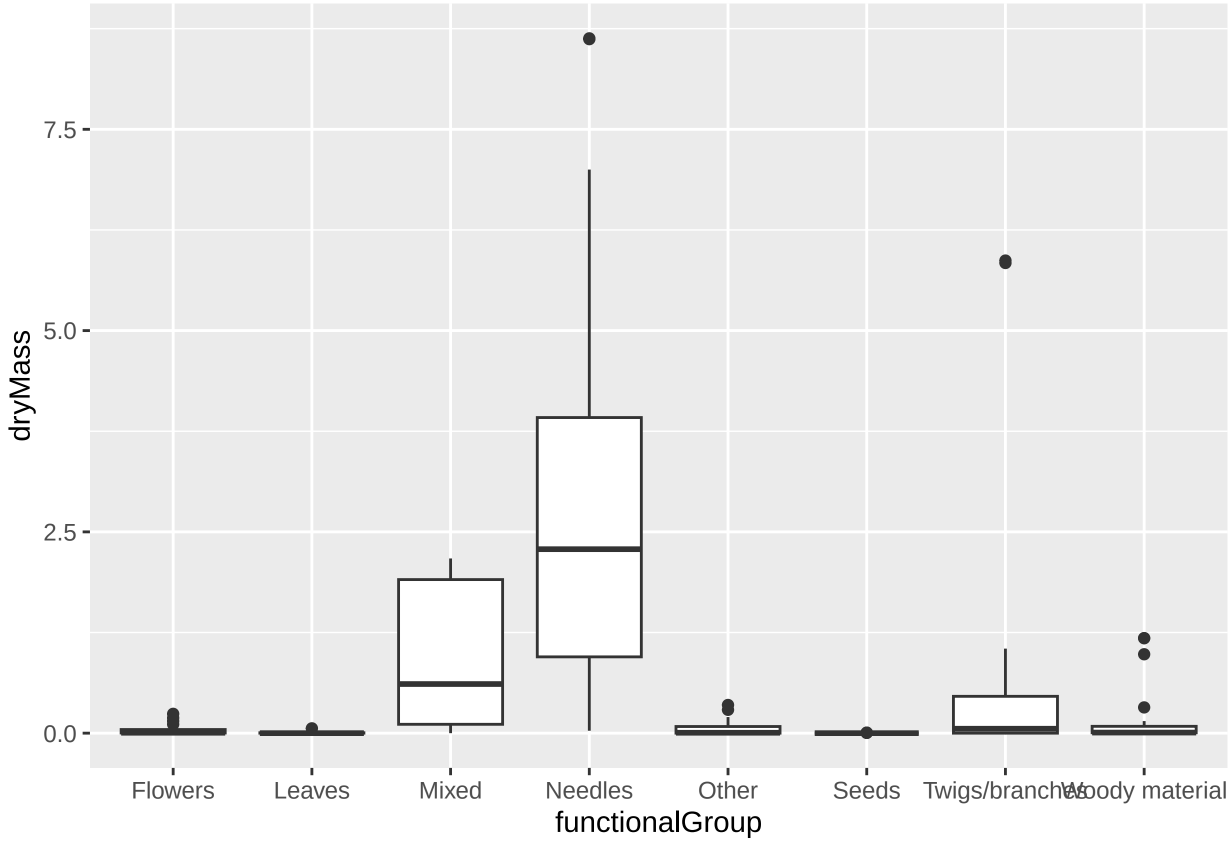
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

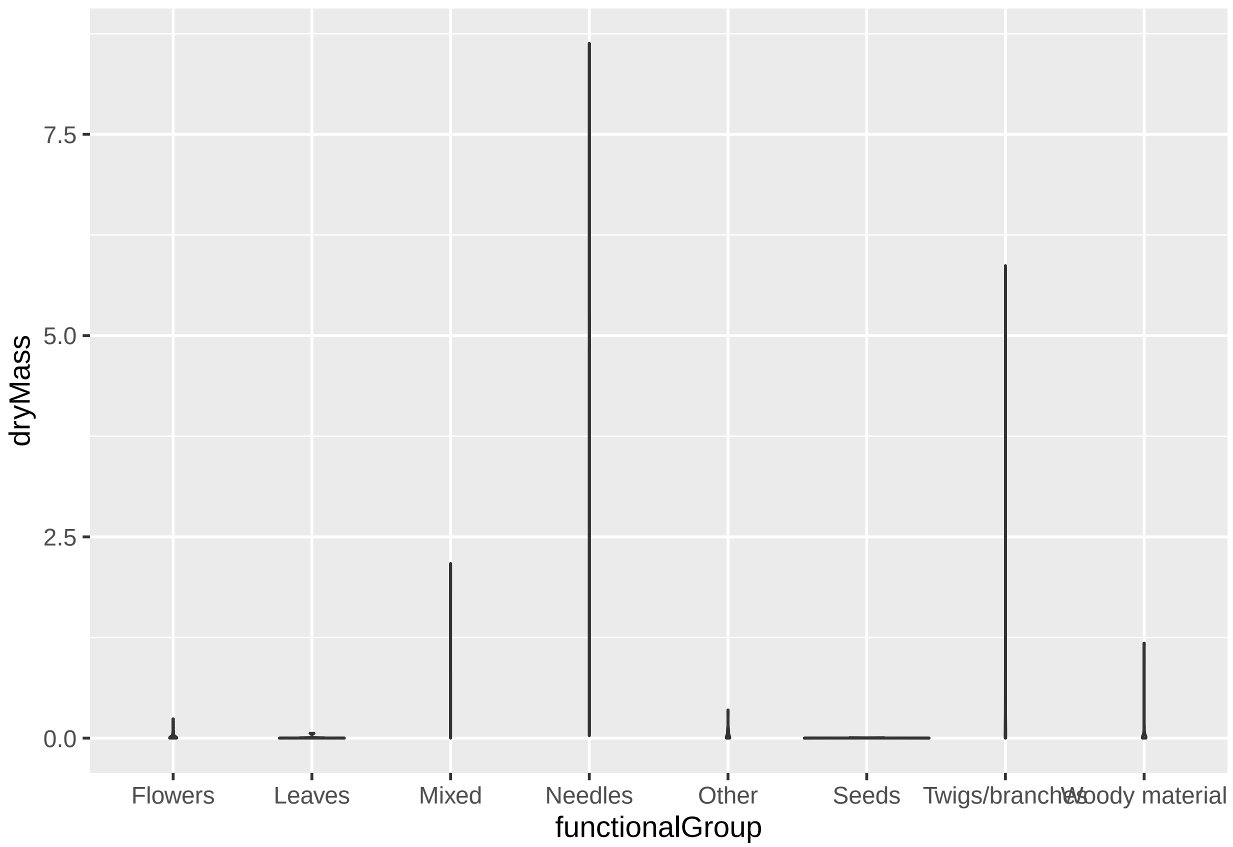


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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



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



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

> Answer: Because there is a fairly equal distribution across the Niwot Ridge sites, a violin plot, which is most helpful when visualizing complex, skewed distributions, would simply show a symmetric shape, something the boxplot already demonstrates. In this case, the violin plots appear as simple vertical and horizontal lines, whereas the boxplots convey more information by clearly showing the median, range, and quartiles of the data.

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

> Answer: Needles and mixed.