Assignment 3: Data Exploration

Li Jia Go

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. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.
6. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai.

The completed exercise is due on Sept 30th.

Set up your R session

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

```
#set up working directory
#getwd()

setwd("/home/guest/R/EDA-Fall2022/")

#load tidyverse
library(tidyverse)

#install.packages('formatR')
#ensure code does not run off page
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=80),
                      tidy=TRUE)

#load and name Neonics dataset, convert strings to factor
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",
                    stringsAsFactors=TRUE)

#load and name Litter dataset, convert strings to factor
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: The widespread use of neonicotinoids in agriculture is toxic to beneficial insects. They include pollinators of vegetables, fruits, flowers and other plants. The enormous rise in the toxicity in America's agricultural lands corresponds to a sharp decline in bees, butterflies, moths and beetles, amongst many other pollinators. Given that such insects pollinate almost 1/3 of all food crops, declining insect numbers can have catastrophic ecological repurcussions. Understanding the ecotoxicology of neonicotinoids can help to provide the crucial scientific evidence required by policy-makers and the public to prevent continued use of the insecticide in agriculture, and prevent greater environmental degradation.

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: In forest ecosystems, litter and woody debris falls is an important component of the nutrient cycle that regulates the accumulation of soil organic matter (SOM), carbon budgets, nutrient cycles and replenishment as well as other ecosystem functions. It can give clues about the rate of forest decay, provide habitat for terrestrial and aquatic organism, as well as influencing water flows and sediment transport. Studying litter and woody debris could potentially provide some insight into the forest ecosystem function and resilience especially in a changing climate.

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 from elevated and ground traps respectively. 2. Ground traps are sampled once per year. Target sampling frequency for elevated traps varies by vegetation present at the site, with frequent sampling (1x every 2weeks) in deciduous forest sites during senescence, and infrequent year-round sampling (1x every 1-2 months) at evergreen sites. 3. litter is defined as material that is dropped from the forest canopy and has a butt end diameter <2cm and a length <50 cm; Fine wood debris is defined as material that is dropped from the forest canopy and has a butt end diameter <2cm and a length >50 cm.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
Neonics_dim <- dim(Neonics)  #dimensions of Neonics
Neonics_dim  #return dimensions i.e. rows, columns
```

```
## [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?

```
Neonics_effect <- summary(Neonics$Effect)
Neonics_effect  #return summary of most common effects
```

```
##     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 common effects studied are population and mortality. These could help to better understand the effect of Neonicotinoids on the population dynamics of various insect 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.

```
# obtaining top 6 most commonly studied species in dataset, the maxsum=7
# (instead of 6) as it would give the top 6 species as well as 1 more output
# combining the rest of the species into the 'other' category
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 6 most commonly studied species in the dataset are as follows (ranked in order): 1. Honey bee 2. Parasitic wasp 3. Buff tailed bumblebee 4. Carniolan honey bee 5. Bumble bee 6. Italian honeybee. They are all beneficial insects, contributing to pollination (bees) or can help to control insect pest infestations (parasitic wasp).

8. Concentrations are always a numeric value. What is the class of Conc.1..Author. in the dataset, and why is it not numeric?

```
class_conc1author <- class(Neonics$Conc.1..Author.)
class_conc1author  #return class of Conc.1..Author.
```

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

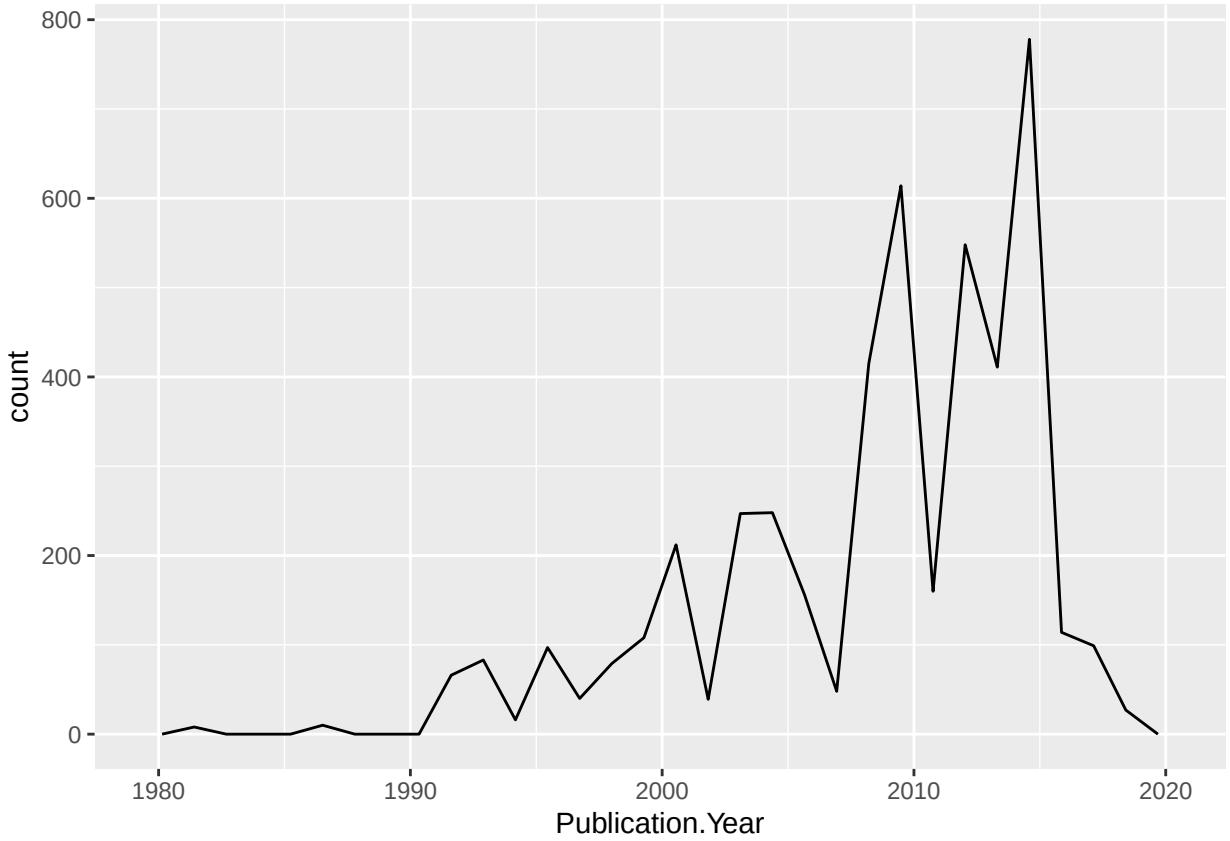
Answer: It is factor data. There are some non-numeric values in the Conc.1..Author. column in the data set which is causing R to read it as factor data (e.g. some numbers in the dataset are succeeded by forward slashes, other non-numeric values in row 67 where "NR" is displayed instead of a number).

Explore your data graphically (Neonics)

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

```
# generating plot of number of studies per publication year
ggplot(Neonics, aes(x = Publication.Year)) + geom_freqpoly()
```

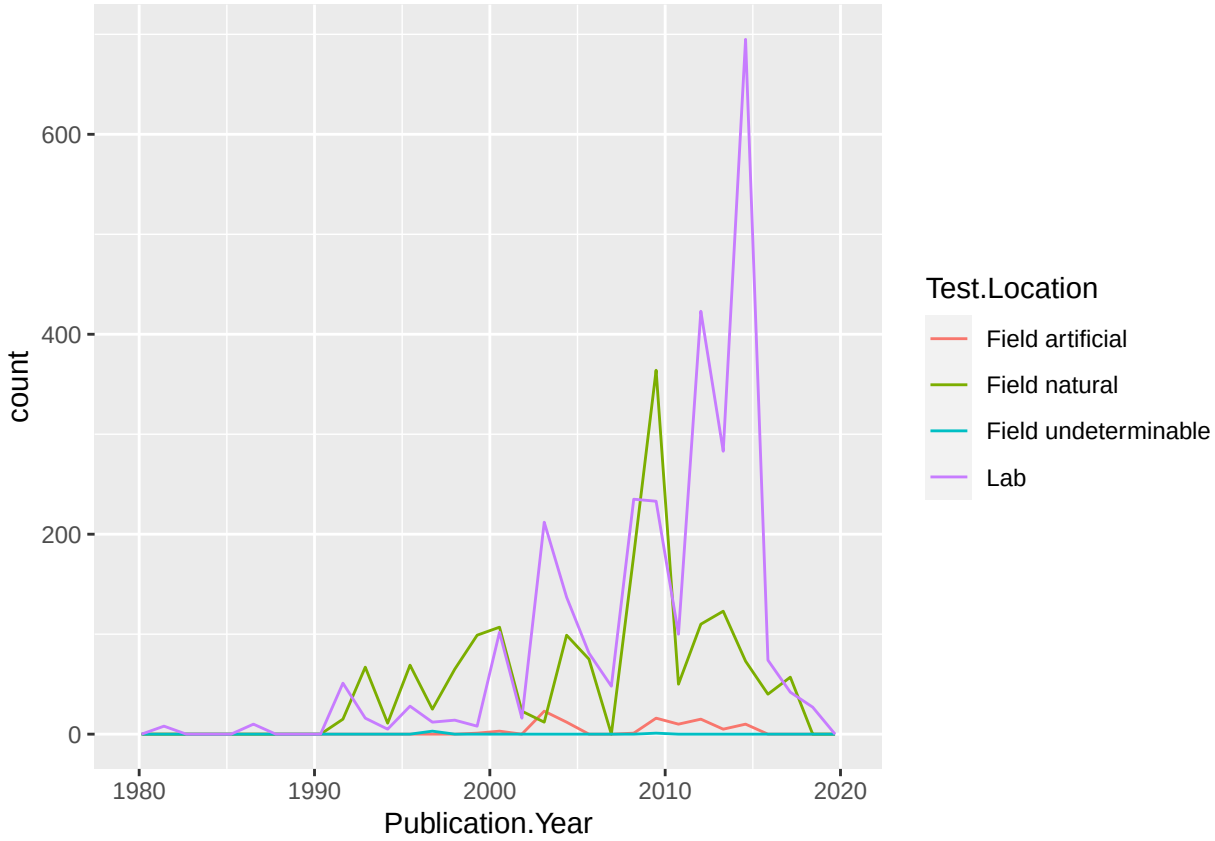
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
# generating plot of studies coloured by categorical data of Test.Location
ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year, color = Test.Location))
```

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

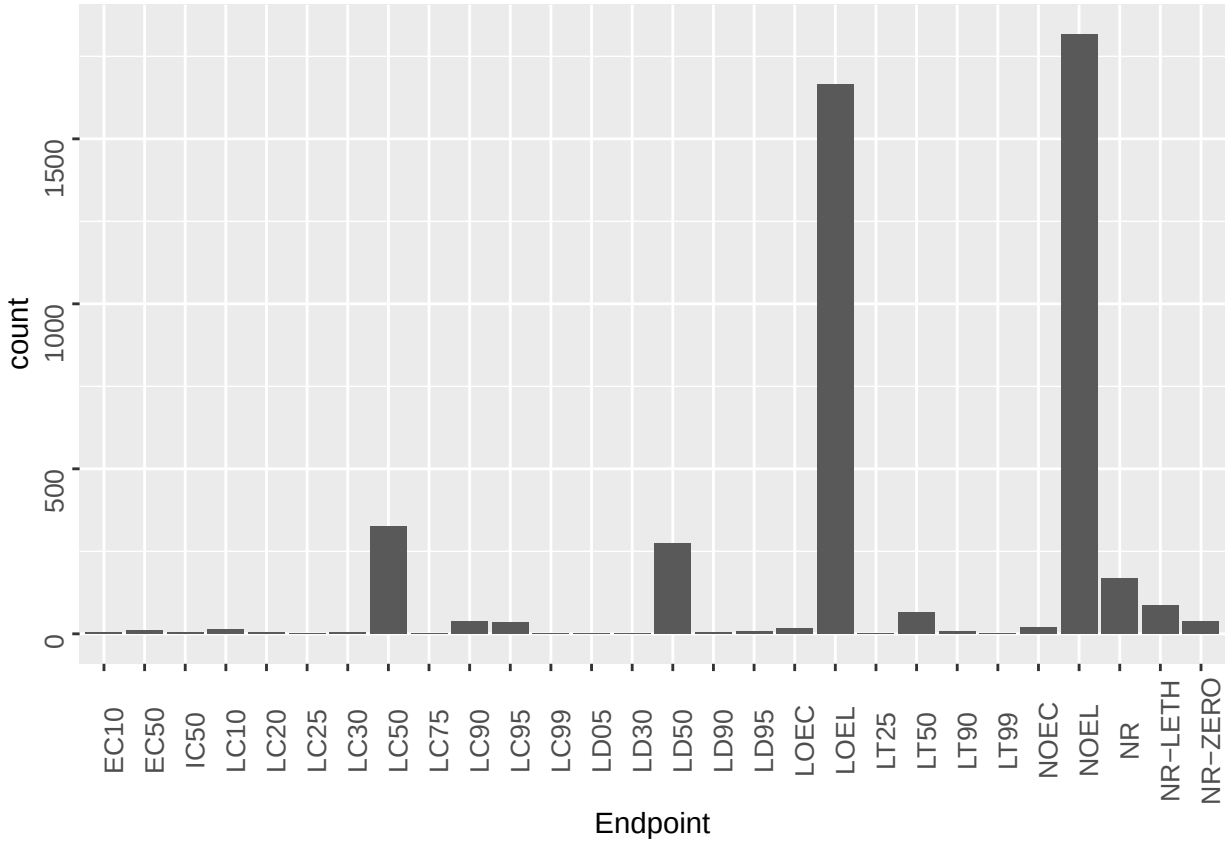


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: Generally, the most common test locations were labs, although the test locations did seem to differ over time. In the early 1990s, there seemed to be a similar proportion of studies using labs or natural field test locations, although there was a slightly larger proportion of studies using natural field test locations in the mid 1990s to 2000. From early 2000-2005, there was a larger proportion of studies using labs, while the late 2000s saw more studies conducted in the natural field. From 2010-2019, lab studies dominated, as compared to natural field studies.

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.

```
# generating bar graph of Endpoint counts and naming it Endpoint_plot
Endpoint_plot <- ggplot(Neonics) + geom_bar(aes(x = Endpoint))

# changing font size to 10 and rotating axis text so that it does not overlap
# with each other
Endpoint_plot + theme(axis.text = element_text(size = 10, angle = 90))
```



Answer: The 2 most common end points are LOEL and NOEL. A toxic endpoint is the result of a study conducted to determine how dangerous a substance is. LOEL refers to the lowest observable effect level, where the lowest dose/concentration of the neonicotinoids produced effects that were significantly different from the responses of controls, as reported by the authors. NOEL refers to no observable effect level, where the highest dose/concentration of neonicotinoids produced effects that were not significantly different from responses of controls, as reported by the authors.

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.

```
# checking if class of collectDate is a date (it is not)
class_collectDate <- class(Litter$collectDate)
class_collectDate
```

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

```
# transforming collectDate into a date
Litter$collectDate <- as.Date(Litter$collectDate)
class_collectDate2 <- class(Litter$collectDate)
class_collectDate2  #confirming the transformation
```

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

```
dates_sampled <- unique(Litter$collectDate)
dates_sampled  #dates sampled were the 2nd Aug 2018 and 30th Aug 2018
```

```
## [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`?

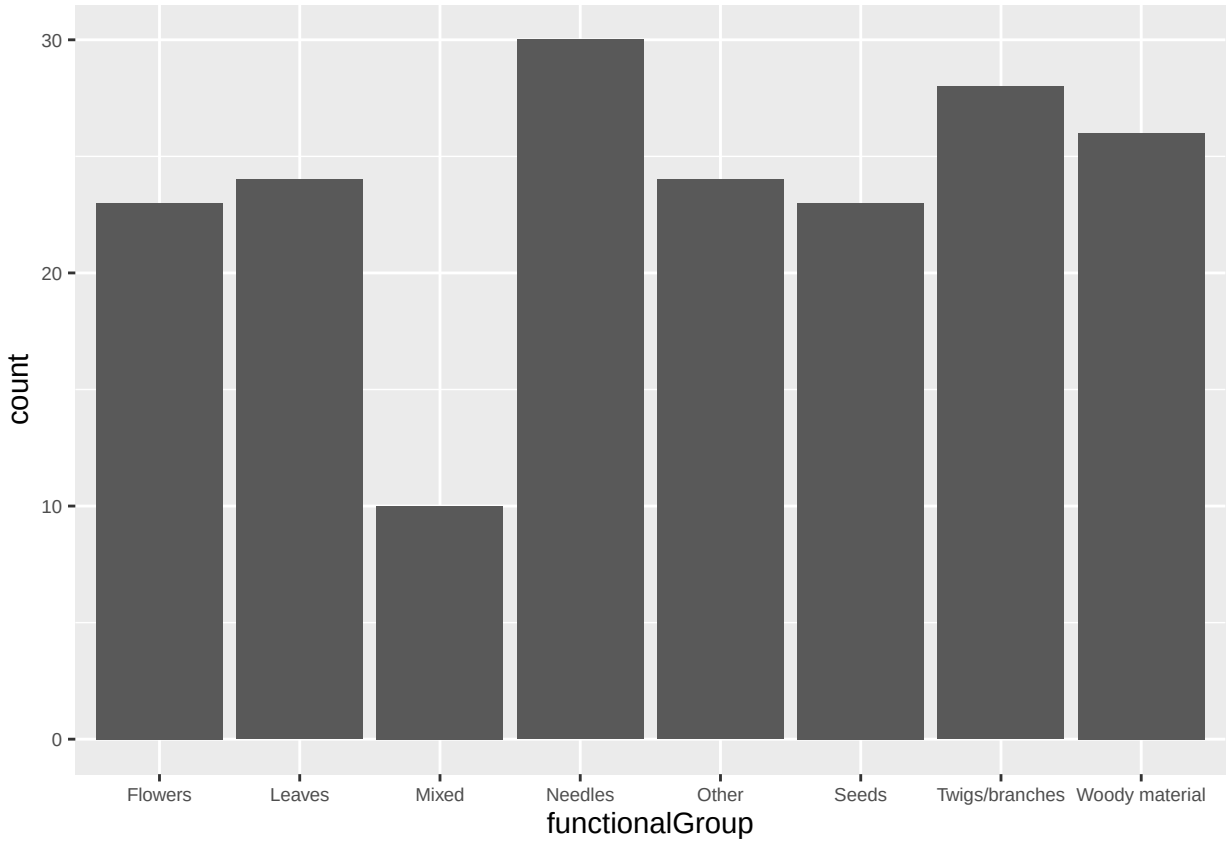
```
number_plotssampled <- (unique(Litter$plotID))
number_plotssampled  #return a list of unique plotIDs
```

```
##  [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: There were 12 unique plots sampled at Niwot Ridge. unique() returns the number of unique plots sampled, as well as the list of their IDs. summary () only returns a list of unique, together with counts of how many times each plot was used in a study.

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.

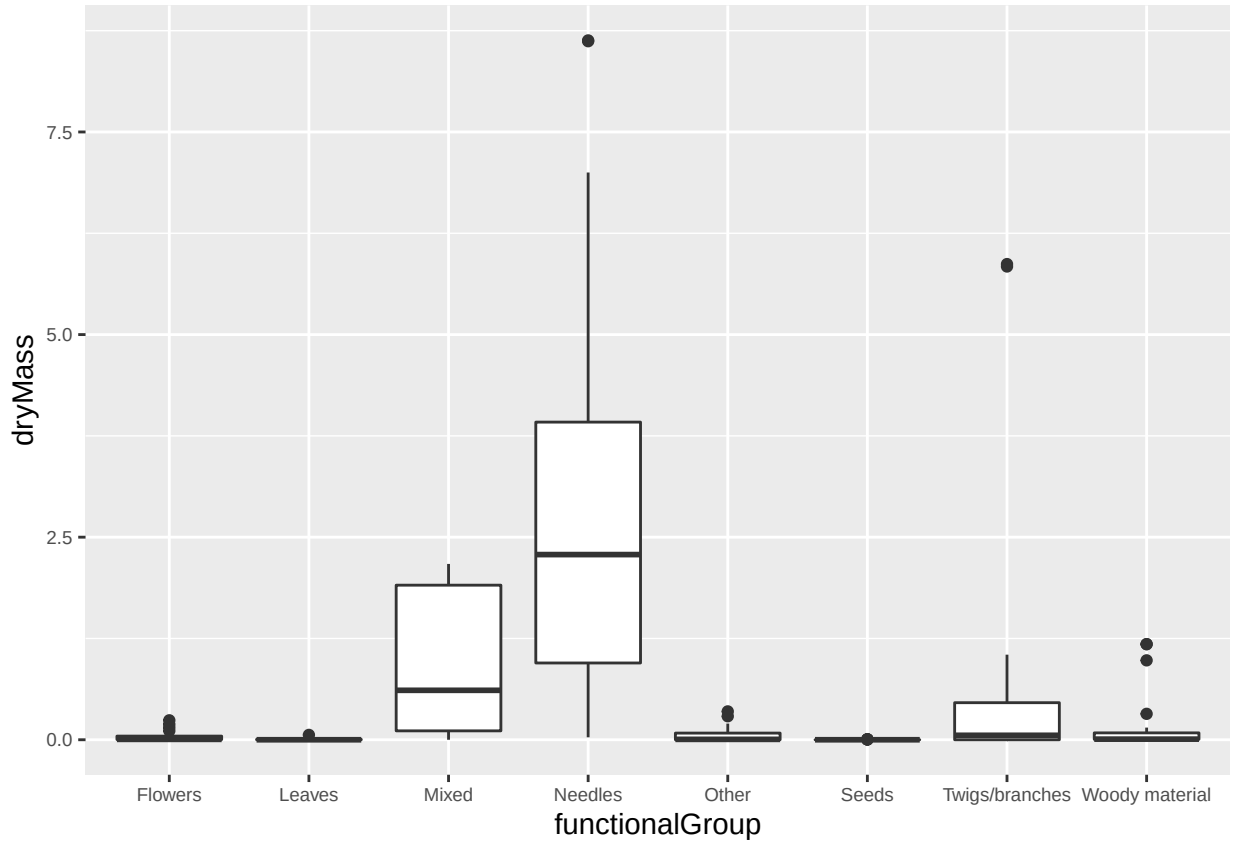
```
# creating bar graph of functionalGroup counts
functionalGroup_plot <- ggplot(Litter, aes(x = functionalGroup)) + geom_bar()

# changing font size to 7 so that axis text does not overlap with each other
functionalGroup_plot + theme(axis.text = element_text(size = 7))
```



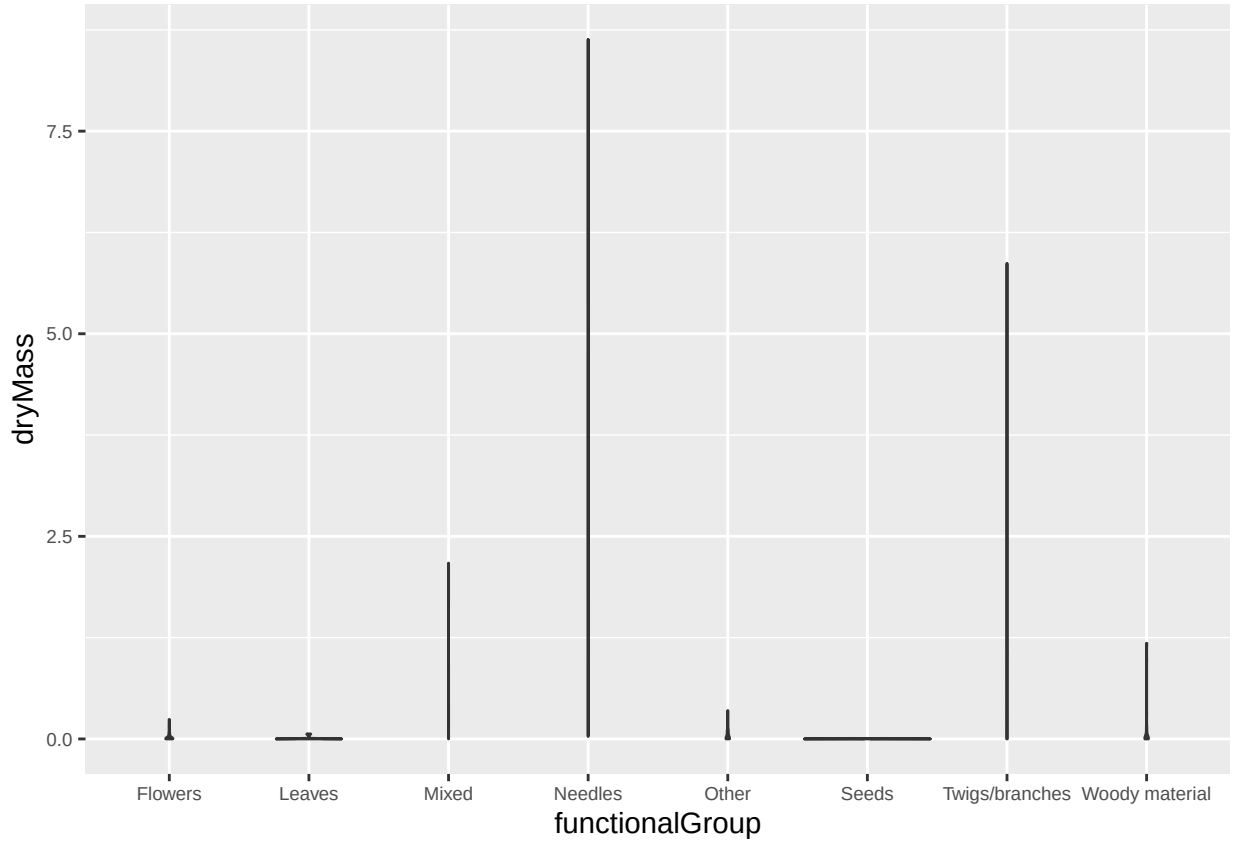
15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
# creating boxplot of dryMass by functionalGroup
dryMass_functionalGroup_boxplot <- ggplot(Litter) + geom_boxplot(aes(x = functionalGroup,
    y = dryMass))
```

```
# changing font size to 7 so that axis text does not overlap with each other
dryMass_functionalGroup_boxplot + theme(axis.text = element_text(size = 7))
```




```
# creating violin plot of dryMass by functionalGroup
dryMass_functionalGroup_violinplot <- ggplot(Litter) + geom_violin(aes(x = functionalGroup,
    y = dryMass))
# changing font size to 7 so that axis text does not overlap with each other
dryMass_functionalGroup_violinplot + theme(axis.text = element_text(size = 7))
```



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

> Answer: There are not enough datapoints in each functional group to display a meaningful density distribution for the violin plot, as such they appear as lines on the plot. The boxplot on the otherhand allows for visualisation of the distribution of a smaller data set based on the minimum, first quartile, median, third quartile and the maximum.

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

> Answer: Needles tend to have the highest biomass at these sites.