Assignment 3: Data Exploration

Joanna Huertas

Spring 2023

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() #Checking working directory
```

```
## [1] "C:/Users/joann/Documents/EDA-Spring2023"
```

```
library(tidyverse) #loading
library(lubridate) #loading
#assigning names to datasets
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: Neonicotinoids can be persistent in the environment and can affect non-target insects, such as pollinators, through exposure to contaminated nectar, pollen, and guttation droplets. These insecticides have been linked to declines in bee populations and other beneficial insects, and reductions in overall biodiversity. As a result, their use has been restricted in some countries. Understanding the toxicity of these insecticides on insects can help in making informed decisions regarding their use and management, and if they should be allowed or not, balancing the benefits of crop protection with minimizing potential harm to the environment and biodiversity.

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:By studying the composition, quantity, and decomposition rate of litter and woody debris, we can gain insights into the functioning of the forest ecosystem, including its productivity and resilience to disturbance. Litter and woody debris provide important habitats for many species of plants, insects, and fungi, and also contribute to the formation of soils.

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. Ground traps are sampled once per year. 3. Trap placement within plots may be either targeted or randomized, depending on the vegetation.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

```
#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?

```
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
```

```
#Neonics$Effect
```

Answer: Population is the most common studied effect, followed by Mortality.Measuring the population of neonicotinoids is important because it can provide information about the exposure levels of these insecticides in the environment. Knowing the population and mortality also helps to monitor the persistence of these insecticides in the environment.

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...]

```
#I found useful to use `summary` and `sort` functions
sorted_data <- sort(summary(Neonics$Species.Common.Name), decreasing = TRUE)

# Extract the first six values
top_six <- sorted_data[1:7]

# Print the top six values
print(top_six)
```

```
##               (Other)             Honey Bee        Parasitic Wasp
##                   670                   667                   285
## Buff Tailed Bumblebee   Carniolan Honey Bee            Bumble Bee
##                   183                   152                   140
##      Italian Honeybee
##                   113
```

Answer:Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, Italian Honeybee are the six most commonly studied species. They are all insects that belong to the Hymenoptera order, which is a diverse group of insects that includes wasps, bees, and ants. They are also all important pollinators, playing a crucial role in the pollination of crops and wildflowers. These insects are valuable to the ecosystem and their populations have declined in many parts of the world due to factors such as habitat loss, pesticide exposure, and disease. As pollinators, they are important for food security and the maintenance of biodiversity.

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"
```

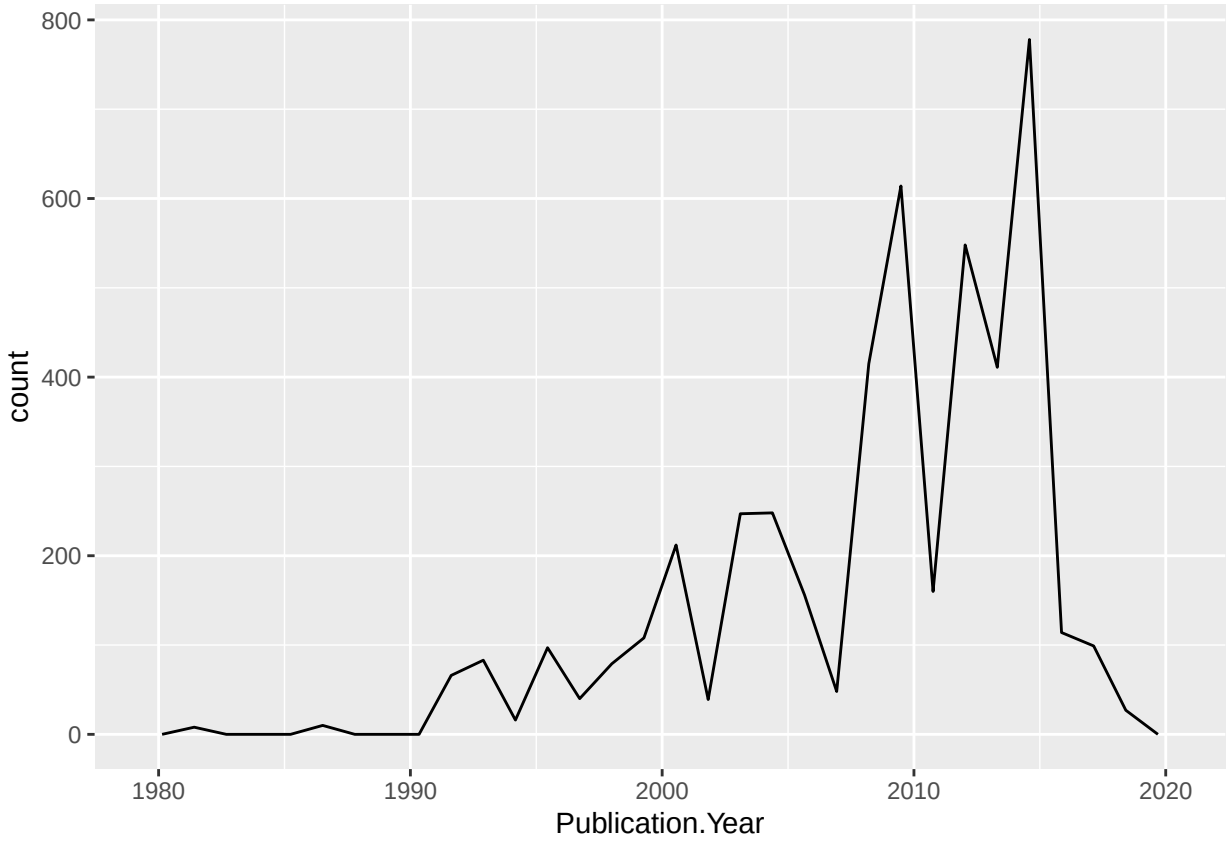
Answer: It is factor. `Conc.1..Author.`is not numeric because some of the values contain symbols (/) and belong to different categories, and not all the concentration was measured in the same way to be compared.

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))
```

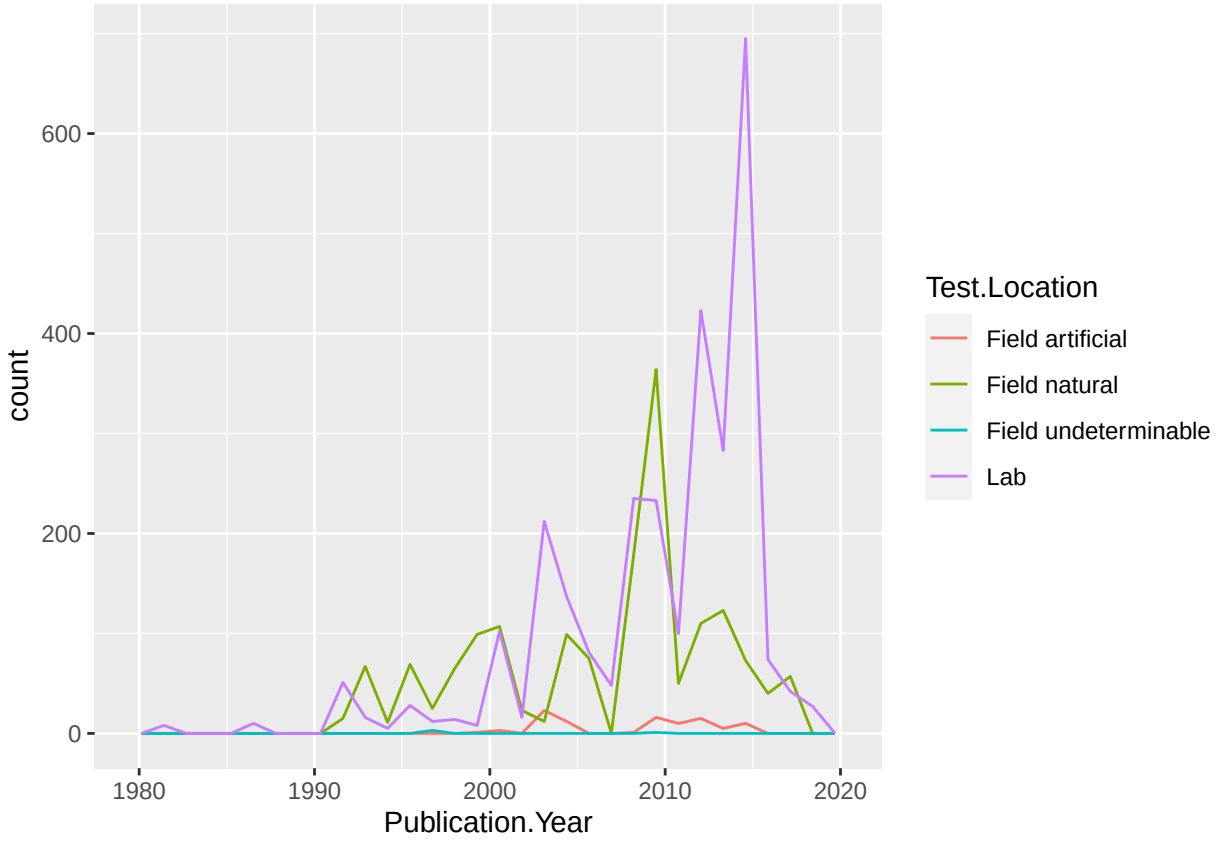
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
ggplot(Neonics)+
  geom_freqpoly(aes(x=Publication.Year, color=Test.Location))
```

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



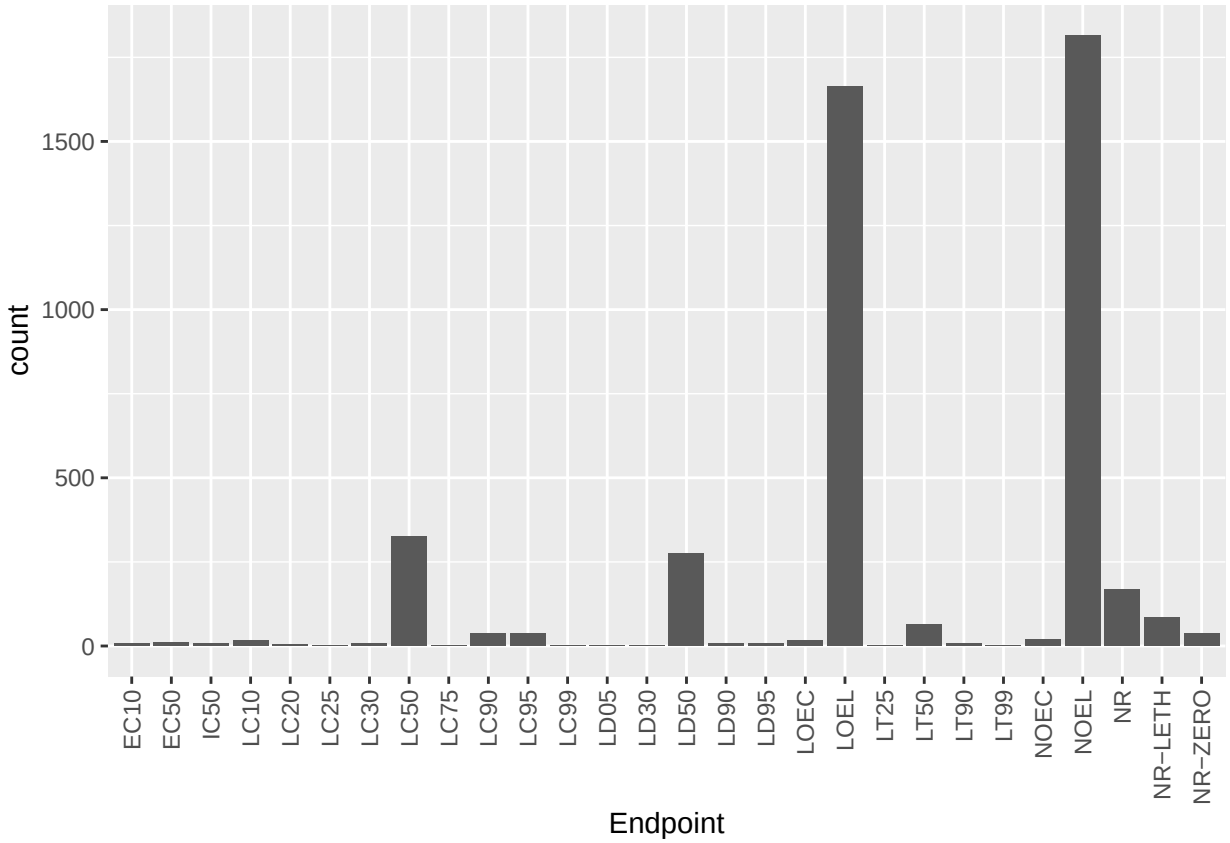
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer:The most common test location is the Lab, followed by field natural. Over time, Lab has been increasing, reaching the maximum number of publications in 2015 approximately and then decreased. The other common location, field natural, reached its maximum before the publication year 2010, and decreased after that.

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(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 the Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC). While 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).Both of them terrestrial.

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)#it is a factor
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate) #changing to date
class(Litter$collectDate)#Now, it is a date
```

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

```
unique(Litter$collectDate) #determining which dates litter was sampled in August 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`?

```
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
```

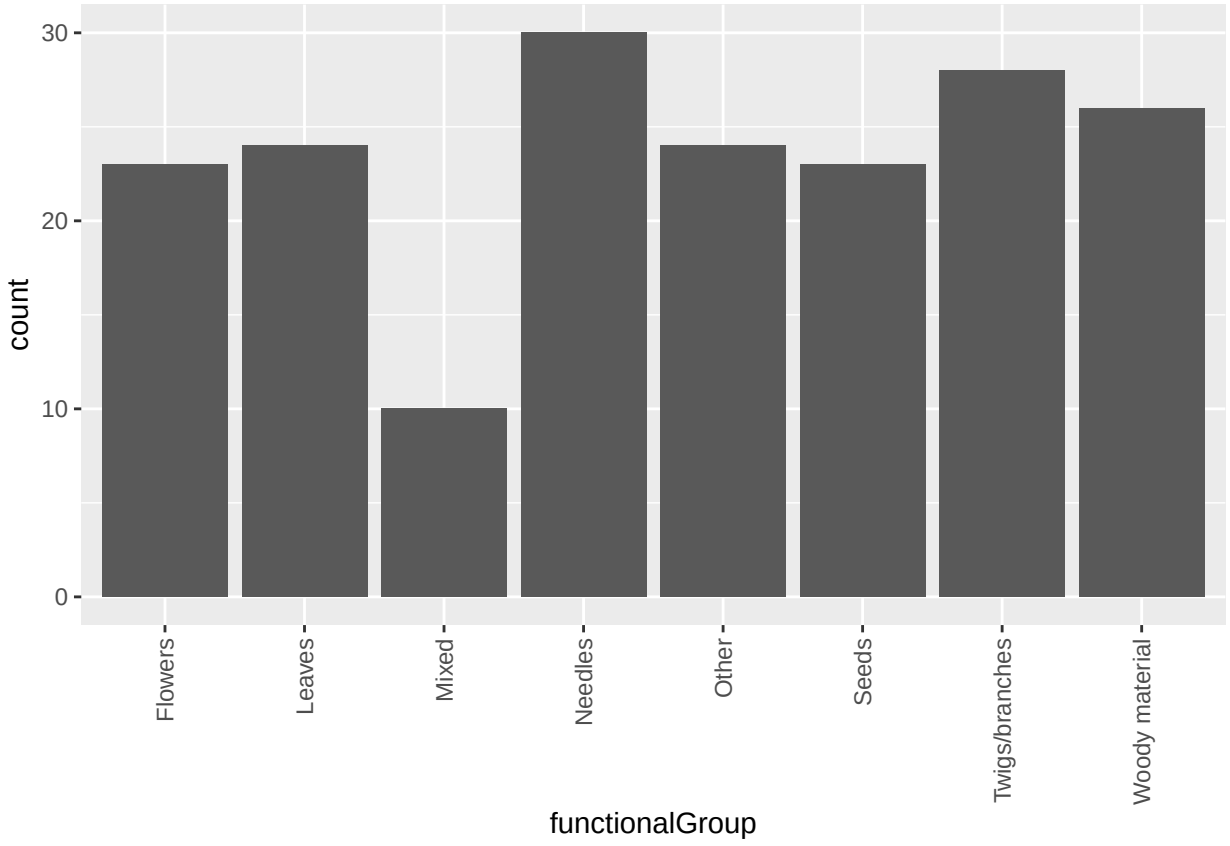
```
summary(Litter$plotID)
```

```
## 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
```

Answer:12 plots were sampled at Niwot Ridge. What we obtain in unique is the number of "levels" or different plots, while with summary we get how many samples we got for each unique plotID.

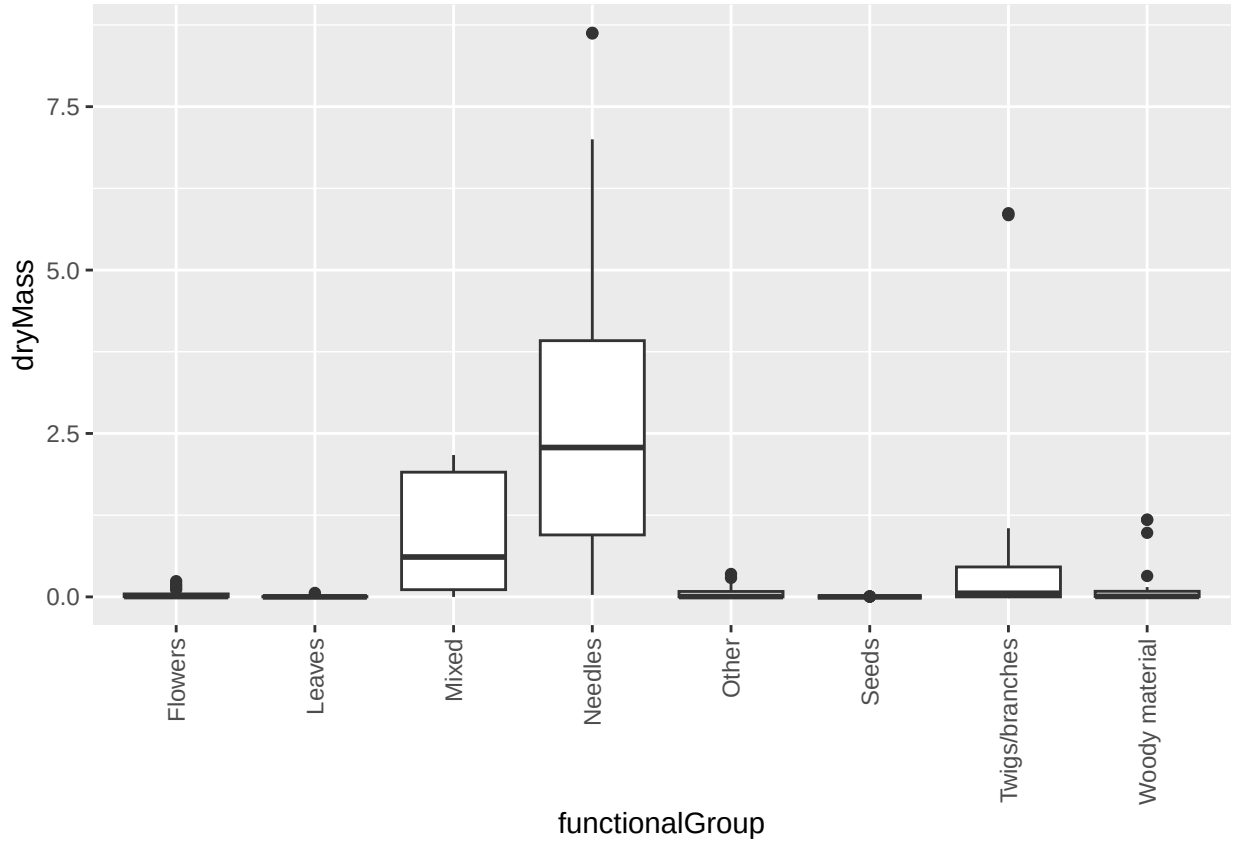
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

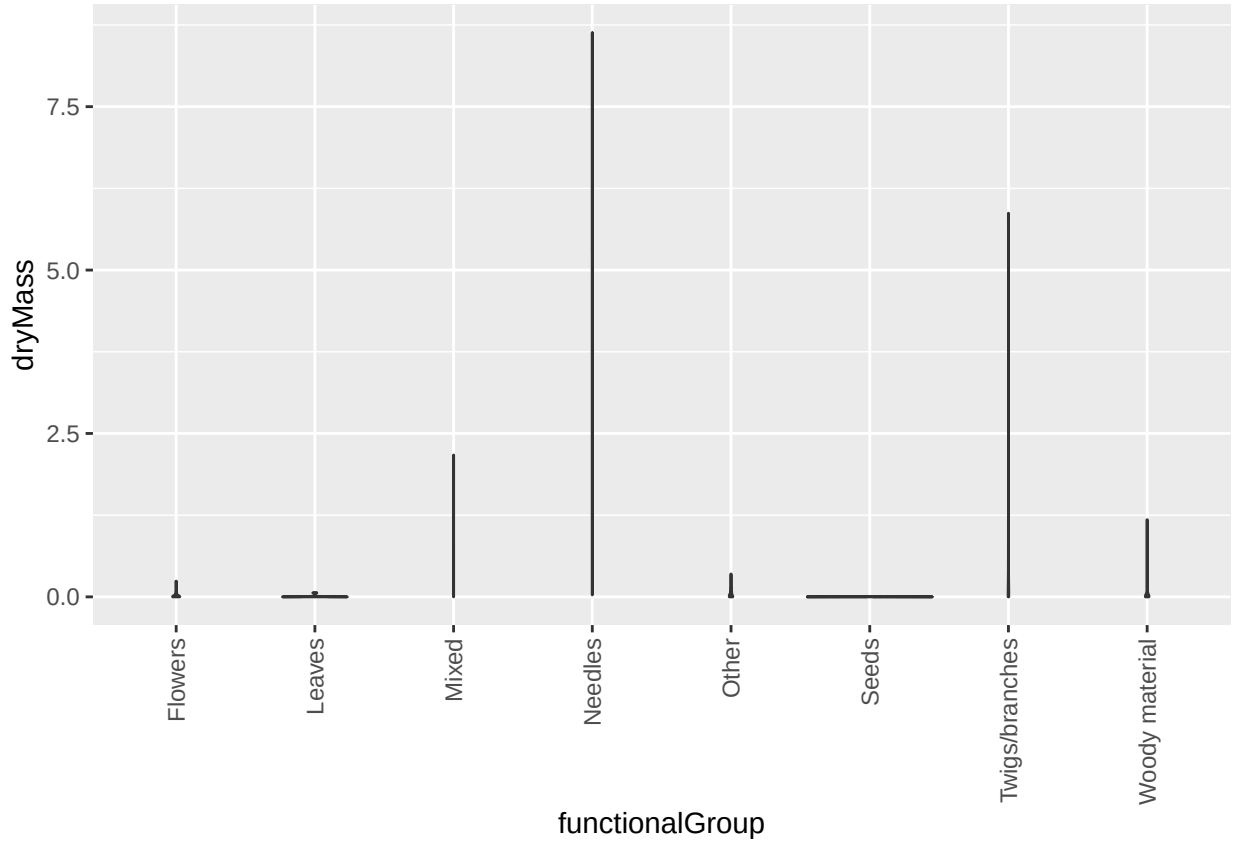


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




```
#
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: We have a relatively small number of observations for each functional group in our data set,and the violin plot shows mostly only a vertical line, which is usually indicative that the data being plotted has very limited variance. In some cases it only shows an horizontal line, and this could be due to either a small data set or a data set where all the values are close to a single value. On the other hand, in the box plots we have, it is easier to visualize the median, upper and lower quartile and outliers. In cases of limited range and variance, the boxplot provides a clearer and more straightforward representation of the central tendency and spread of the data.

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

> Answer:Needles,looking at the boxplot.