Assignment 3: Data Exploration

Jessalyn Chuang

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 packages for tidyverse, lubridate, and here
library(tidyverse)
library(lubridate)
library(here)
#Command points to the project directory
here()
```

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

```r
#upload Neonics dataset
Neonics <- read.csv(
  file = here("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"),
  stringsAsFactors = TRUE)

#upload Litter dataset
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 Knowledge base, 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: Researching the ecotoxicology of neonicotinoids on insects is important due to their detrimential impacts on pollinator populations, which are extremely important for ecosystem health and agriculture. These insecticides can disrupt beneficial insect roles, harm biodiversity, and have negative impacts on human health. Understanding their effects can help researchers provide insights that inform policies around the pesticide's usage. Ultimately, this research can play a pivotal role in promoting sustainable agricultural practices, safeguarding public helath, and preserving biodiversity. (NRDC, 2022)

   Natural Resources Defense Council. (2002, May 25). Neonicotinoids 101: The effects on humans and bees. NRDC. https://www.nrdc.org/stories/neonicotinoids-101-effects-humans-and-bees

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 in forests is important for understanding ecosystem dynamics. These components play key roles in nutrient cycling, soil health, and carbon sequestration, contributing to forest productivity and resilience. They provide habitat and food for diverse organisms, influence fire behavior, and affect water retention and quality. Additionally, monitoring litter and debris over time helps assess ecological changes due to climate change and land-use practices, informing adaptive management strategies for forest conservation.

   Scheungrab, M. J., & Olson, R. L. (2002). Silvicultural practices in the South to achieve sustainability of wildlife habitats and diversity (General Technical Report SRS-038). U.S. Department of Agriculture, Forest Service, Southern Research Station. https://www.srs.fs.usda.gov/pubs/gtr/gtr_srs038/gtr_srs038-scheungrab001.pdf

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 is collected using elevated traps while fine woody debris is collected in ground traps. 2. Litter is sorted into categories like leaves, twigs, and seeds, and dry weights are measured with 0.01 grams precision. 3. Ground traps are sampled annually, while elevated traps are sampled biweekly in deciduous forests and monthly/bimonthly in evergreen forests.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#Returns dimensions of Neonics which is 4623 rows by 30 columns
dim(Neonics)
```

```
## [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...]

```
#Returns the most common effects of Neonics studied
sort(summary(Neonics$Effect), decreasing = TRUE)
```

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

Answer: The most common effects studied are population and mortality. These effects are interesting necause they help wit hstudying the ecological impacts of these pesticides. Studying population effects helps us understand how exposure might impact species reproduction, while mortality studies may reveal direct toxic effects that could lead to population decline.

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 top 7 most frequently studied species were selected from
#Species.Common.Name because maxsum aggregates the remaining species under
#"Other" and includes it in the ranking.
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 top six most commonly studied include: (1) Honey Bee (2) Parasitic Wasp (3) Buff Tailed Bumblebee (4) Carniolan Honey Bee (5) Bumble Bee (6) Italian Honeybee. Honey bees, bumble bees, and their subspecies are pollinators. All of these species are pollinators and might be of interest over other insects because they are crucial to crop pollinaton, and as a result, global food production. They are also important indicators of ecosystem health. The decline of these species may have significant ecological and economic consequences.

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...]

```
#Call upon the class of Conc.1..Author column, which returns "factor"
class(Neonics$Conc.1..Author.)
```

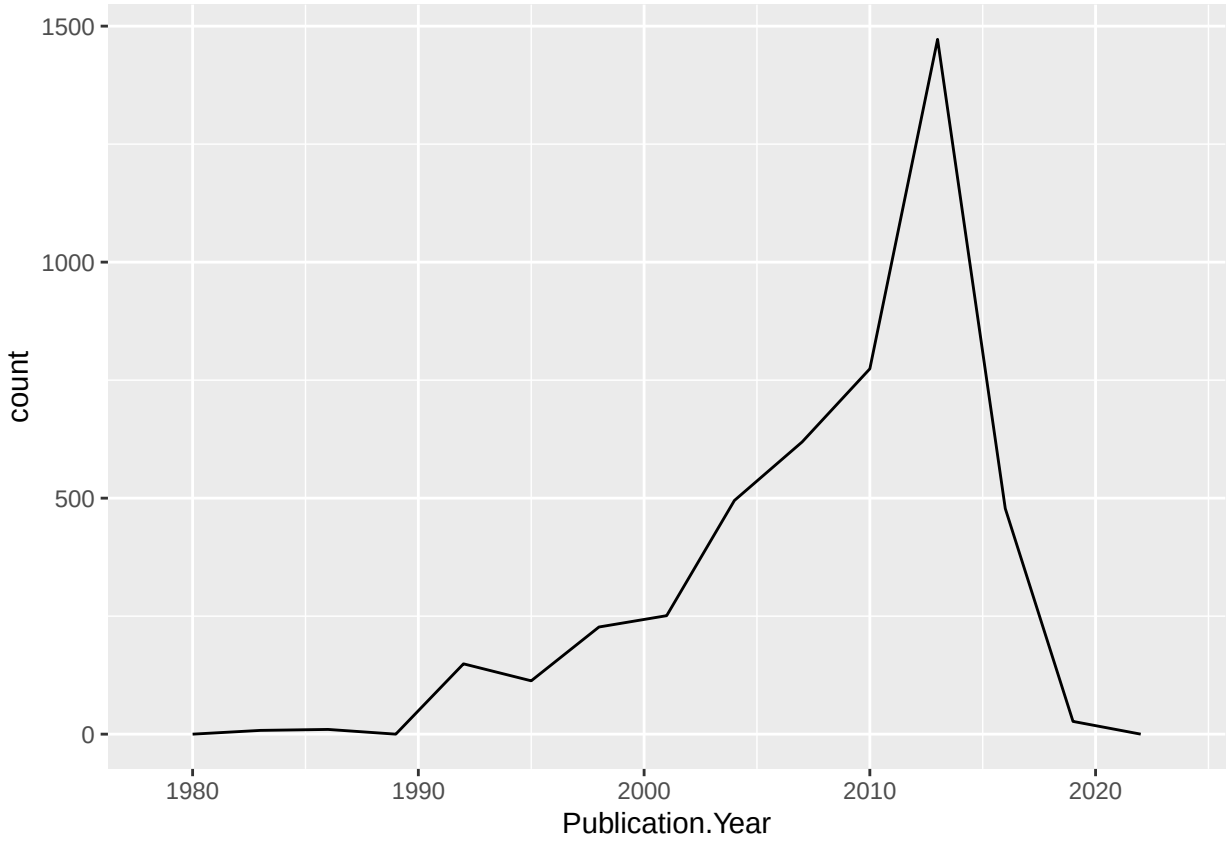
```
## [1] "factor"
```

Answer: R is interpreting the column as a factor instead of numeric. Several of the rows contain non-numeric characters like "NR" or "~" or "/" causing R to treat the entire column as a factor.

Explore your data graphically (Neonics)

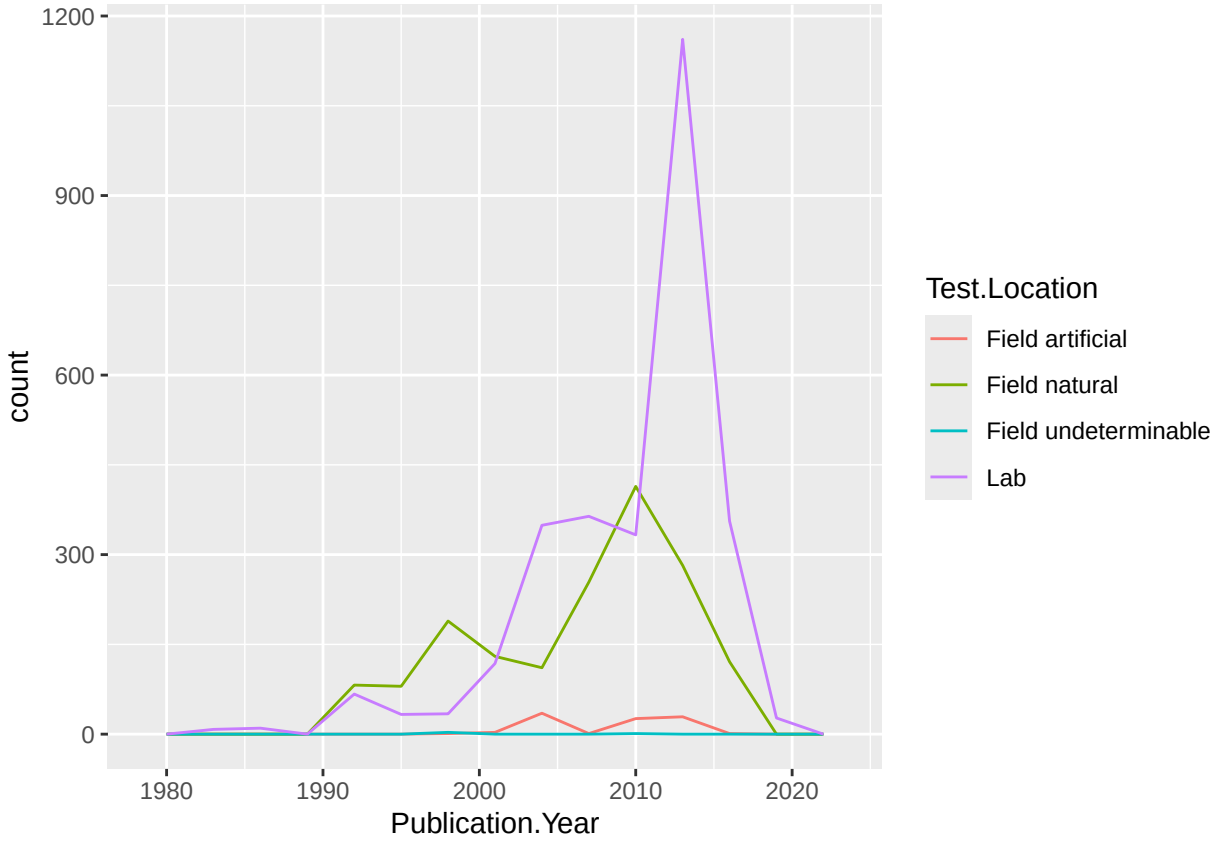
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
#Creation of geom_freqpoly that shows number of studies conducted by
#publication year. Binwidth was set to 3 to be able to smooth out the curve
#helps lessen the visual "spikiness" that comes with years with less data
#or "NR" values
ggplot(Neonics,aes(x=Publication.Year))+
  geom_freqpoly(binwidth = 3)
```



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

```
#Breaks apart the studies per publication year into how many of each type of
#Test.Location in each public year.Binwidth was set to 3 to be able to smooth
#out the curve again.
ggplot(Neonics,aes(x=Publication.Year,color = Test.Location))+
  geom_freqpoly(binwidth = 3)
```



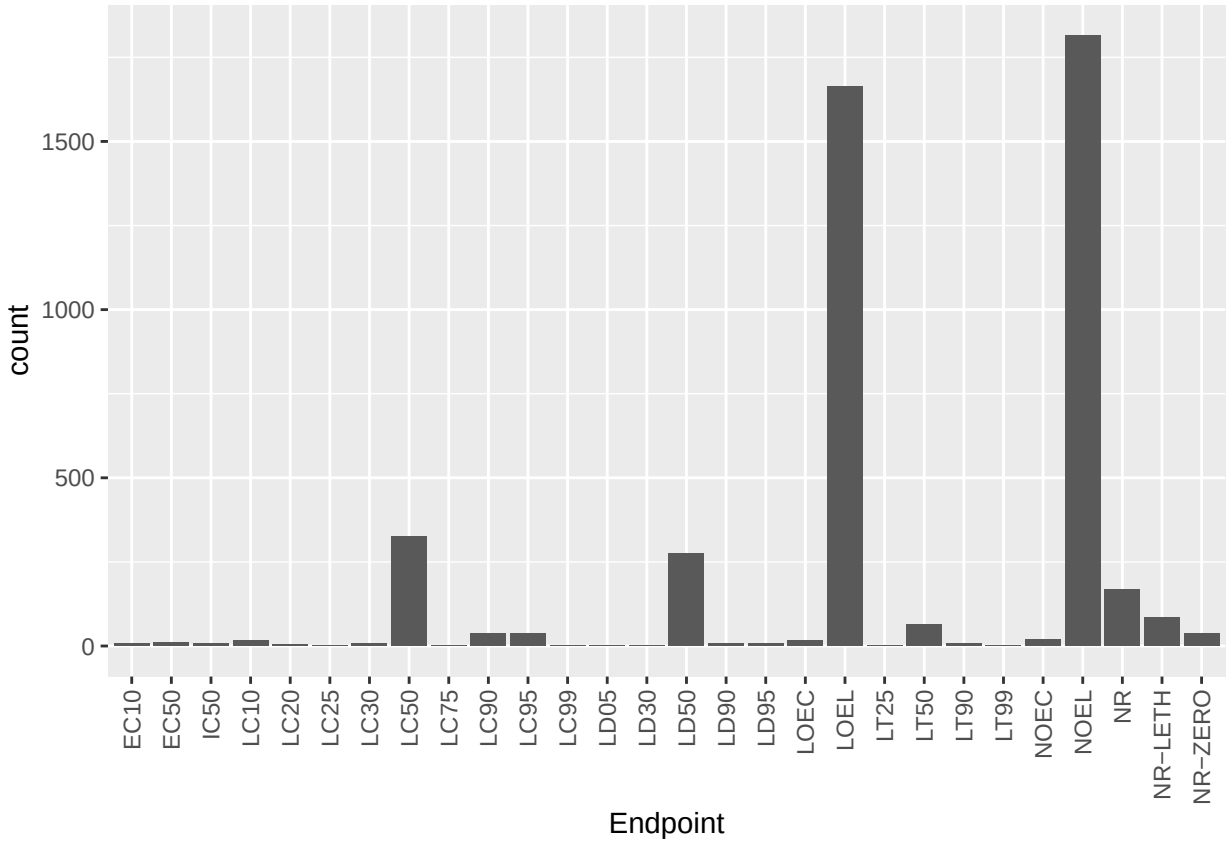
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations seem to be "Lab" and "Field Natural". These two have interchanged in terms of which has the highest count. Between 1990-2000, more of the test locations were in "Field Natural", but this soon switched between 2000-2010, when the Lab Test.Location became more common. Lab also soared above Field Natural during 2010-2020.

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...]

```
#Creation of a geom_bar graph of Endpoint counts
ggplot(Neonics,aes(x=Endpoint))+
  geom_bar()+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: NOEL and LOEL are the two most common end points. NOEL is a Terrestrial term for no-observable-effect-level (the highest concentration dose producing effects not significantly different from responses of controls according to the author's reported statistical test). LOEL is a Terrestrial term for lowest-observable-effect-level (the lowest concentration dose producing effects that were significantly different as reported by authors from responses of controls).

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 before class transformation. It is a "factor" to
#start
class(Litter$collectDate)
```

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

```
#Use of lubridate to change the collectDate column to date type
Litter$collectDate <- ymd(Litter$collectDate)
#new class is "Date"
class(Litter$collectDate)
```

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

```r
#Returns the unique dates that are in the collectDate column
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`?

```r
#Returns the unique plotIDs in Litter
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
```

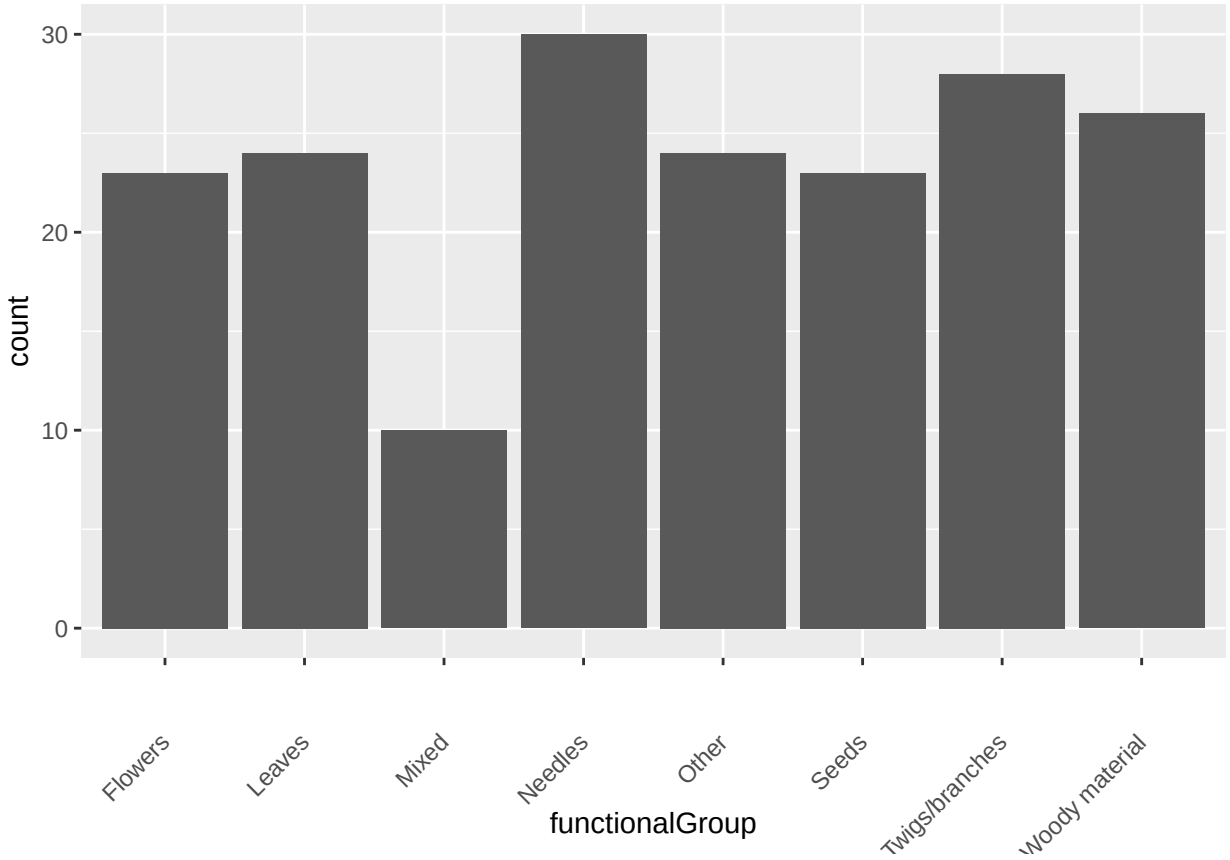
```r
summary(Litter$plotID)
```

```
## 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 different plots were sampled at Niwot Ridge. While unique() shows you the different levels, summary() produces the count that each level shows up.

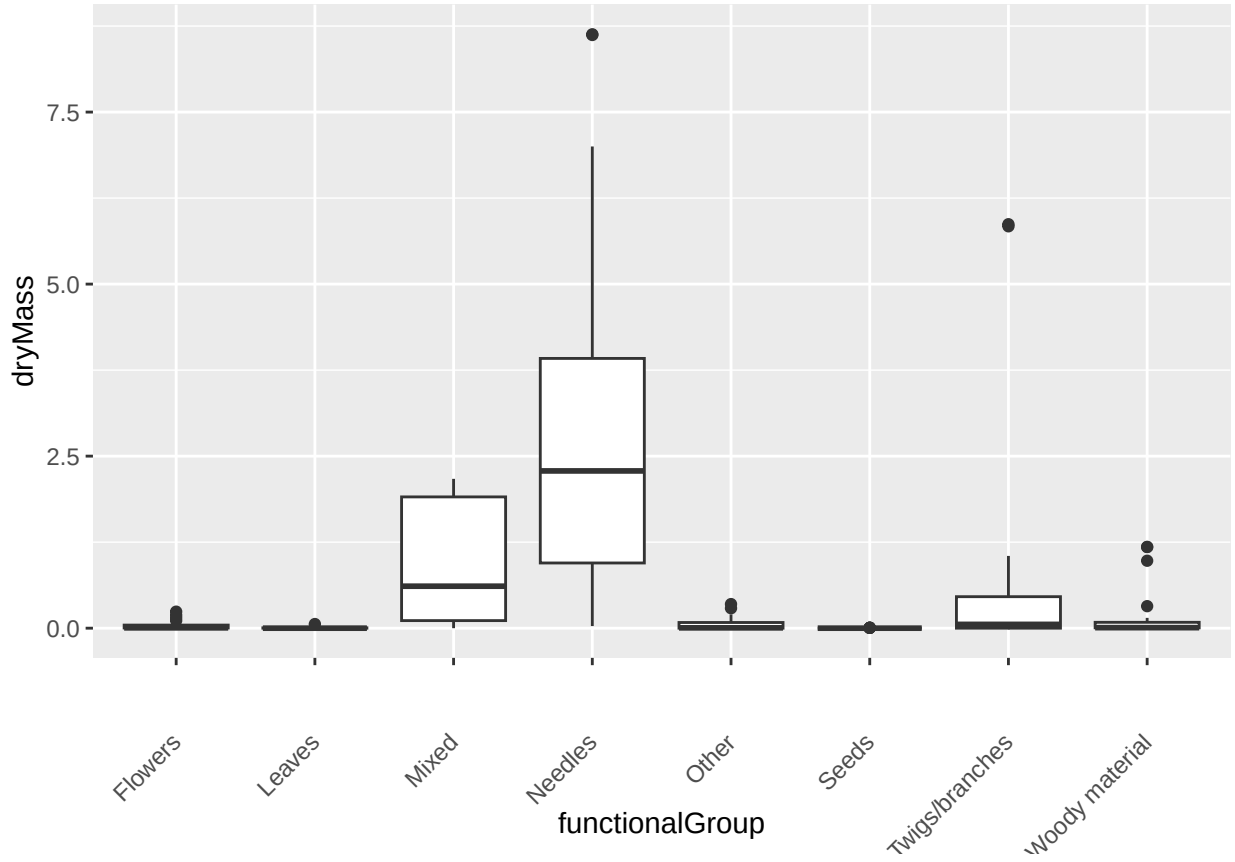
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
#bar graph of functionalGroup counts
ggplot(Litter, aes(x = functionalGroup))+
  geom_bar()+
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust=1))
```

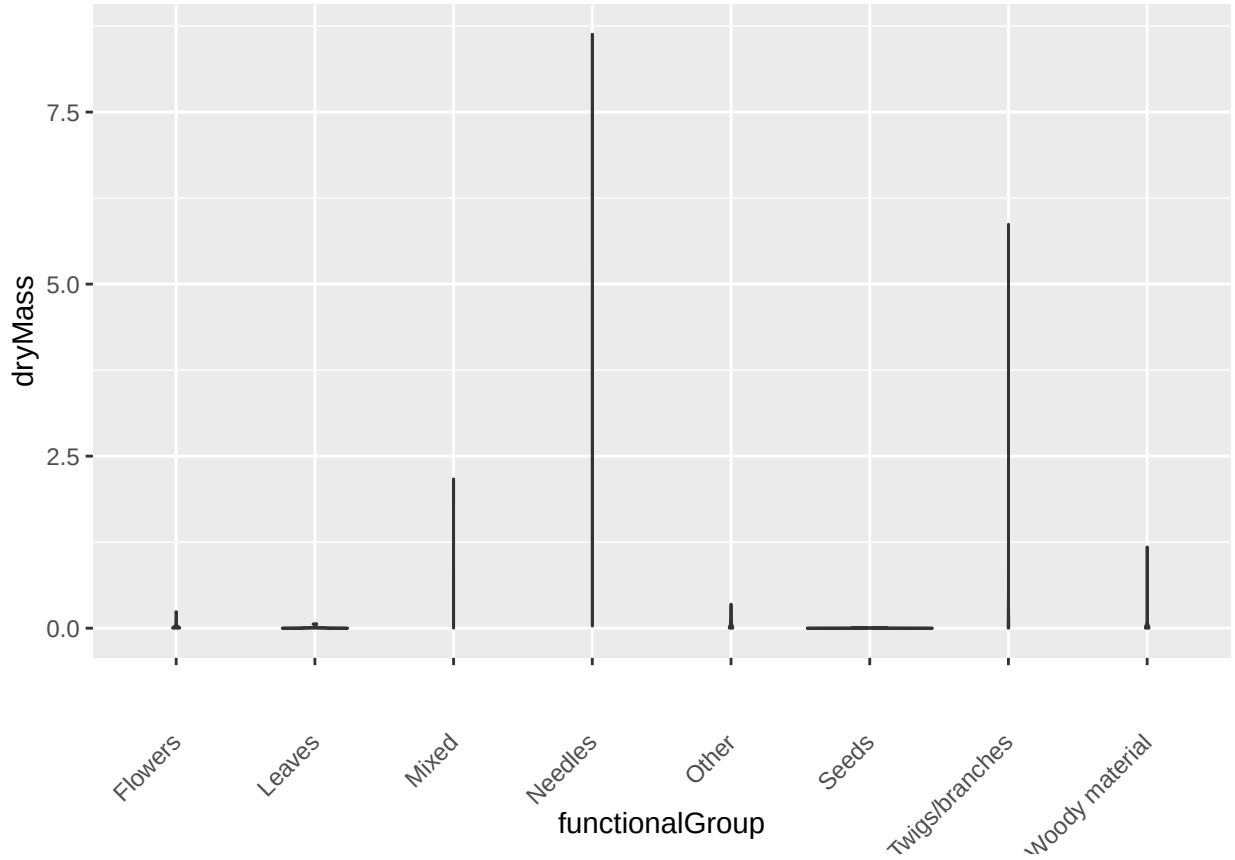


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#box plot creation of dryMass by functionalGroup
ggplot(Litter)+
  geom_boxplot(aes(x = functionalGroup, y = dryMass))+
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust=1))
```




```
#violin plot creation of dryMass by functionalGroup
ggplot(Litter)+
  geom_violin(aes(x = functionalGroup, y = dryMass))+
  theme(axis.text.x = element_text(angle = 45, 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 more effective in this case because it provides a clear summary of key statistical measures (median, quartiles, and potential outliers) without being affected by the small sample size within each functionalGroup. In contrast, the violin plot includes a density estimate that can become misleading when the number of observations is small. Since the width of the violin plot is proportional to the density of values, small sample sizes may result in inaccurate or overly smoothed density shapes, making it harder to interpret the distribution properly. The boxplot, by focusing on summary statistics, offers a cleaner and more accurate representation of the data distribution for small groups.

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

> Answer: Needles have the highest biomass at these sites followed by mixed litter then twigs/branches.