Assignment 3: Data Exploration

Kayla Emerson

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 necessary packages
library(tidyverse)
library(lubridate)
library(here)

#check current working directory
getwd()
```

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

```
#upload ECOTOX neonicotinoid dataset
neonics <- read.csv(
  file = here('Data','Raw','ECOTOX_Neonicotinoids_Insects_raw.csv'),
  stringsAsFactors = T
)

#upload litter and debris dataset
litter <- read.csv(
  file = here('Data','Raw','NEON_NIWO_Litter_massdata_2018-08_raw.csv'),
  stringsAsFactors = T
)
```

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: We are interested in the ecotoxicology of neonicotinoids on insects because they could leave lasting impacts on certain insect species, including pollinators, which are very important for several plants and crops.

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 that falls on the forest ground can give us more insight as to how that ecosystem stores and cycles nutrients and carbon because litter is the first step in decomposition. It can also tell us more about the biodiversity and the health of the ecosystem.

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 was collected into elevated PVC pipes and ground traps 2.Ground traps were sampled once per year, elevated traps were sampled more frequently 3. Mass data for samples were obtained to an accuracy of 0.01 grams

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#get dimensions of neonics

dim(neonics)
```

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

```
length(neonics)
```

```
## [1] 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...]

```
#get summary of the "effect" column of neonics
summary(neonics$Effect)
```

```
##     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
```

```
#sort effect column
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: Behavior, mortality, and population are the most common effects that are studied. Hormones, histology, and physiology are the least common effects studied.

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...]

```
#get summary of the studied species
summary(neonics$Species.Common.Name)
```

```
##                          Honey Bee                     Parasitic Wasp
##                                667                                285
##              Buff Tailed Bumblebee                Carniolan Honey Bee
##                                183                                152
##                         Bumble Bee                   Italian Honeybee
##                                140                                113
##                    Japanese Beetle                  Asian Lady Beetle
##                                 94                                 76
##                     Euonymus Scale                           Wireworm
##                                 75                                 69
##                  European Dark Bee                  Minute Pirate Bug
##                                 66                                 62
##               Asian Citrus Psyllid                      Parastic Wasp
##                                 60                                 58
##             Colorado Potato Beetle                    Parasitoid Wasp
##                                 57                                 51
##                Erythrina Gall Wasp                       Beetle Order
##                                 49                                 47
##        Snout Beetle Family, Weevil           Sevenspotted Lady Beetle
##                                 47                                 46
##                     True Bug Order              Buff-tailed Bumblebee
##                                 45                                 39
##                       Aphid Family                     Cabbage Looper
##                                 38                                 38
##               Sweetpotato Whitefly                      Braconid Wasp
##                                 37                                 33
##                       Cotton Aphid                     Predatory Mite
##                                 33                                 33
##             Ladybird Beetle Family                         Parasitoid
##                                 30                                 30
##                      Scarab Beetle                      Spring Tiphia
##                                 29                                 29
##                        Thrip Order               Ground Beetle Family
##                                 29                                 27
##                 Rove Beetle Family                      Tobacco Aphid
##                                 27                                 27
##                       Chalcid Wasp             Convergent Lady Beetle
##                                 25                                 25
##                      Stingless Bee                  Spider/Mite Class
##                                 25                                 24
##                Tobacco Flea Beetle                   Citrus Leafminer
##                                 24                                 23
##                    Ladybird Beetle                          Mason Bee
##                                 23                                 22
##                           Mosquito                      Argentine Ant
##                                 22                                 21
##                             Beetle         Flatheaded Appletree Borer
##                                 21                                 20
##               Horned Oak Gall Wasp                 Leaf Beetle Family
##                                 20                                 20
##                  Potato Leafhopper         Tooth-necked Fungus Beetle
##                                 20                                 20
##                       Codling Moth          Black-spotted Lady Beetle
##                                 19                                 18
##                       Calico Scale                Fairyfly Parasitoid 
##                                 18                                 18 
##                        Lady Beetle             Minute Parasitic Wasps 
##                                 18                                 18 
##                          Mirid Bug                   Mulberry Pyralid 
##                                 18                                 18 
##                           Silkworm                     Vedalia Beetle 
##                                 18                                 18 
##              Araneoid Spider Order                          Bee Order 
##                                 17                                 17 
##                     Egg Parasitoid                       Insect Class 
##                                 17                                 17 
##           Moth And Butterfly Order       Oystershell Scale Parasitoid 
##                                 17                                 17 
## Hemlock Woolly Adelgid Lady Beetle              Hemlock Wooly Adelgid 
##                                 16                                 16 
##                               Mite                        Onion Thrip 
##                                 16                                 16 
##              Western Flower Thrips                       Corn Earworm 
##                                 15                                 14 
##                  Green Peach Aphid                          House Fly 
##                                 14                                 14 
##                          Ox Beetle                 Red Scale Parasite 
##                                 14                                 14 
##                 Spined Soldier Bug              Armoured Scale Family 
##                                 14                                 13 
##                   Diamondback Moth                      Eulophid Wasp 
##                                 13                                 13 
##                  Monarch Butterfly                      Predatory Bug 
##                                 13                                 13 
##              Yellow Fever Mosquito                Braconid Parasitoid 
##                                 13                                 12 
##                       Common Thrip       Eastern Subterranean Termite 
##                                 12                                 12 
##                             Jassid                         Mite Order 
##                                 12                                 12 
##                          Pea Aphid                   Pond Wolf Spider 
##                                 12                                 12 
##           Spotless Ladybird Beetle             Glasshouse Potato Wasp 
##                                 11                                 10 
##                           Lacewing            Southern House Mosquito 
##                                 10                                 10 
##            Two Spotted Lady Beetle                         Ant Family 
##                                 10                                  9 
##                       Apple Maggot                            (Other) 
##                                  9                                670
```

```
#get top 6 species using maxsum
summary(neonics$Species.Common.Name, maxsum = 6)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee 
##                   667                   285                   183 
##   Carniolan Honey Bee            Bumble Bee               (Other) 
##                   152                   140                  3196
```

Answer: The six most commonly studied species in the dataset are all pollinators - bees and wasps. These species might be of interest over other insects because they help to pollinate other plants and are essential to ecosystems all over the globe.

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…]

```
#get the class of "Conc.1..Author"
class(neonics$Conc.1..Author.)
```

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

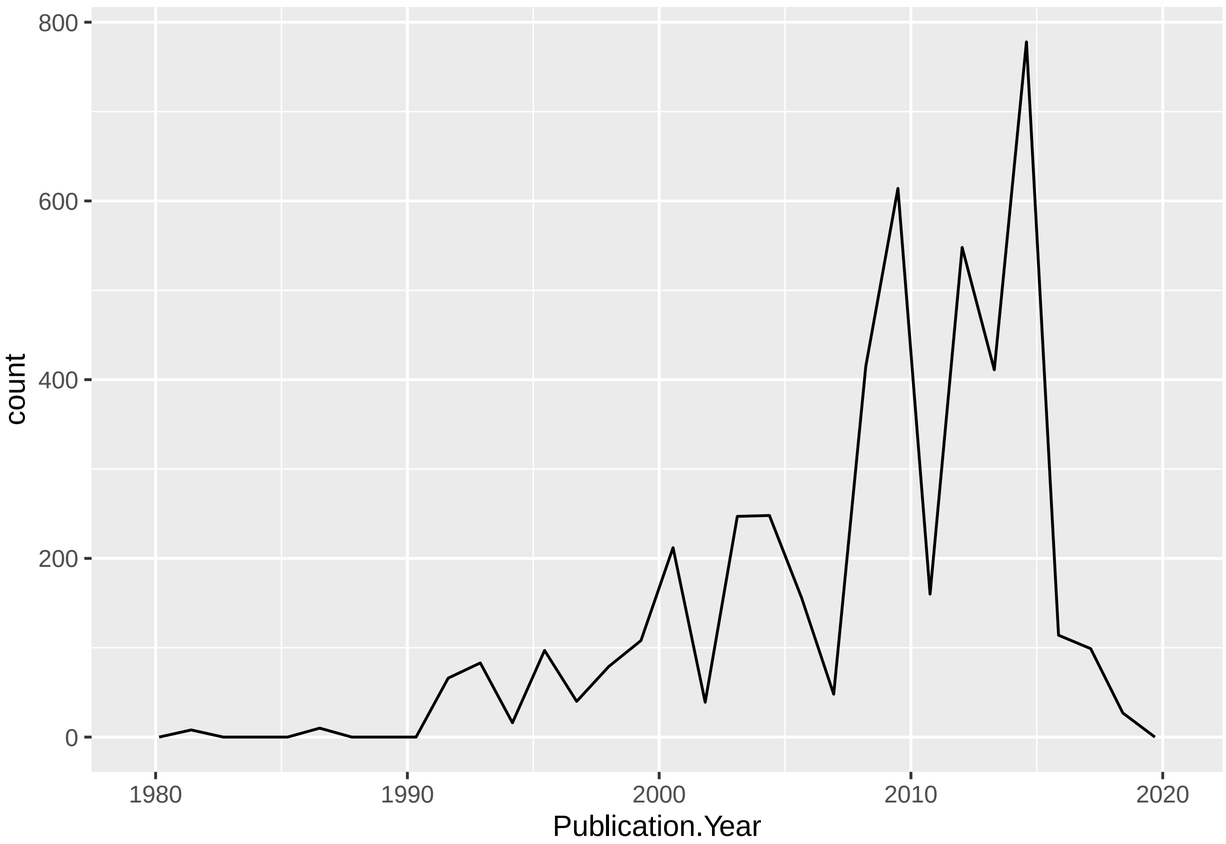
Answer: The class of "Conc.1..Author" column in the dataset is factor because some values were labeled as "NR" or included "/" and we are unsure what that means in terms of numbers.

Explore your data graphically (Neonics)

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

```
#generate a plot of the number of studies conducted by pub. year using geom_freqpoly
ggplot(neonics) +
geom_freqpoly(aes(x = Publication.Year))
```

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



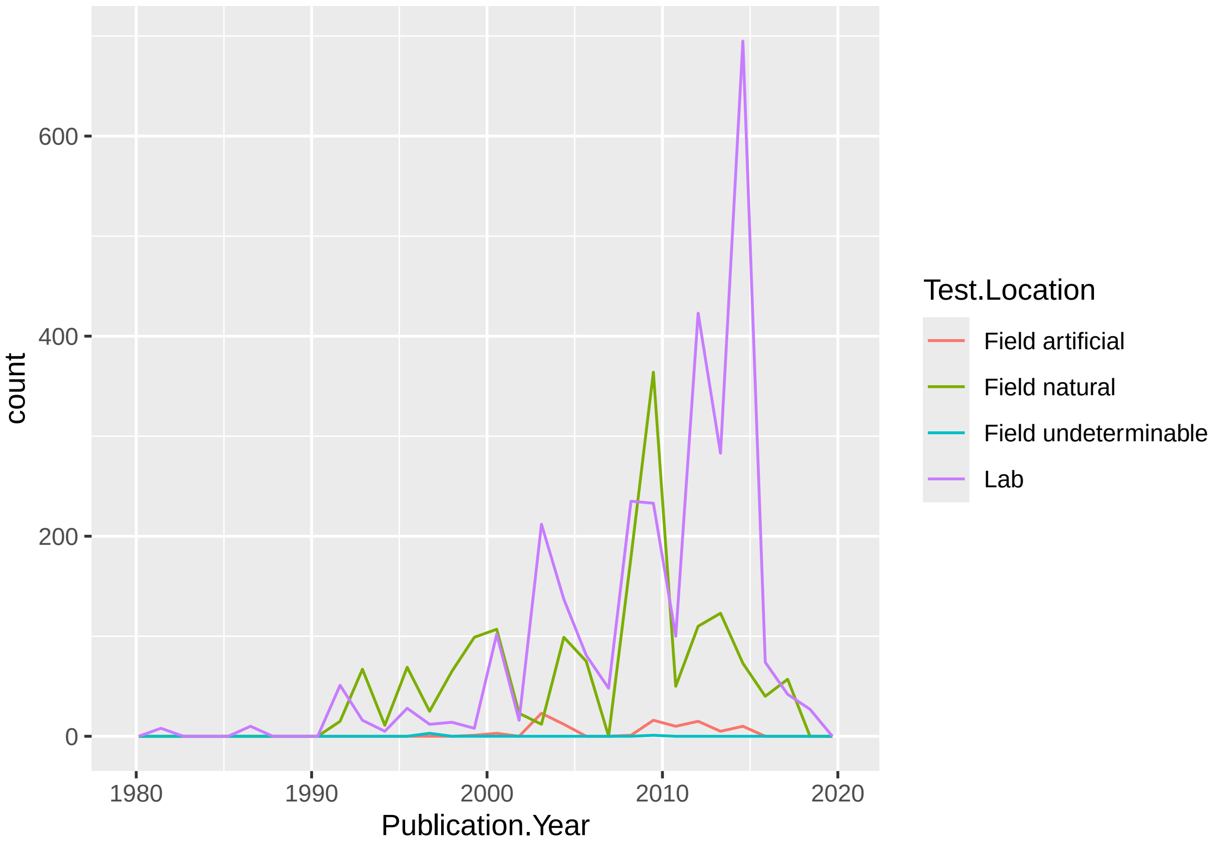
10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
class(neonics$Test.Location)
```

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

```
#Add test location as an additional variable, and categorize by color
ggplot(neonics) +
geom_freqpoly(aes(x = Publication.Year, color = Test.Location))
```

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



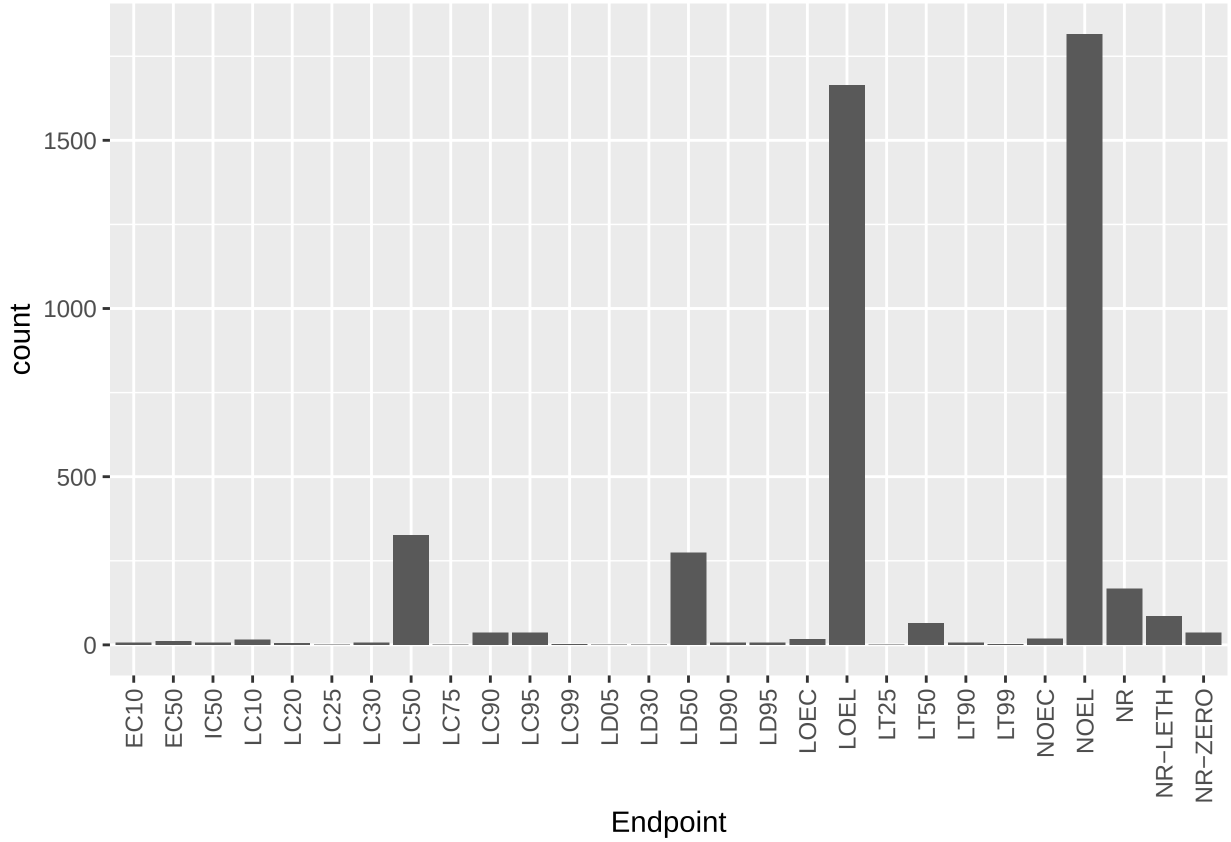
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are “Field natural” and “Lab.” These are the most common over time but way more so since 1990. After 2010, “Lab” was by far the most common test location.

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…]

```r
#Create a bar graph of Endpoint counts
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 end points are LOEL and NOEL. LEOL stands for Lowest-observable-effect-level which means it was the lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls. NOEL stands for No-observable-effect-level which means it's the highest dose (concentration) producing effects not significantly different 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.

```r
#get the class of collectDate
class(litter$collectDate)
```

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

```r
#change collectDate from factor to date
litter$collectDate = ymd(litter$collectDate)

#confirm class change
class(litter$collectDate)
```

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

```r
#amazing, moving on

#use unique function to determine which dates litter was sampled
unique(litter$collectDate)
```

```
## [1] "2018-08-02" "2018-08-30"
```

```r
#only twice...
```

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
#determine different plots sampled using unique
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
```

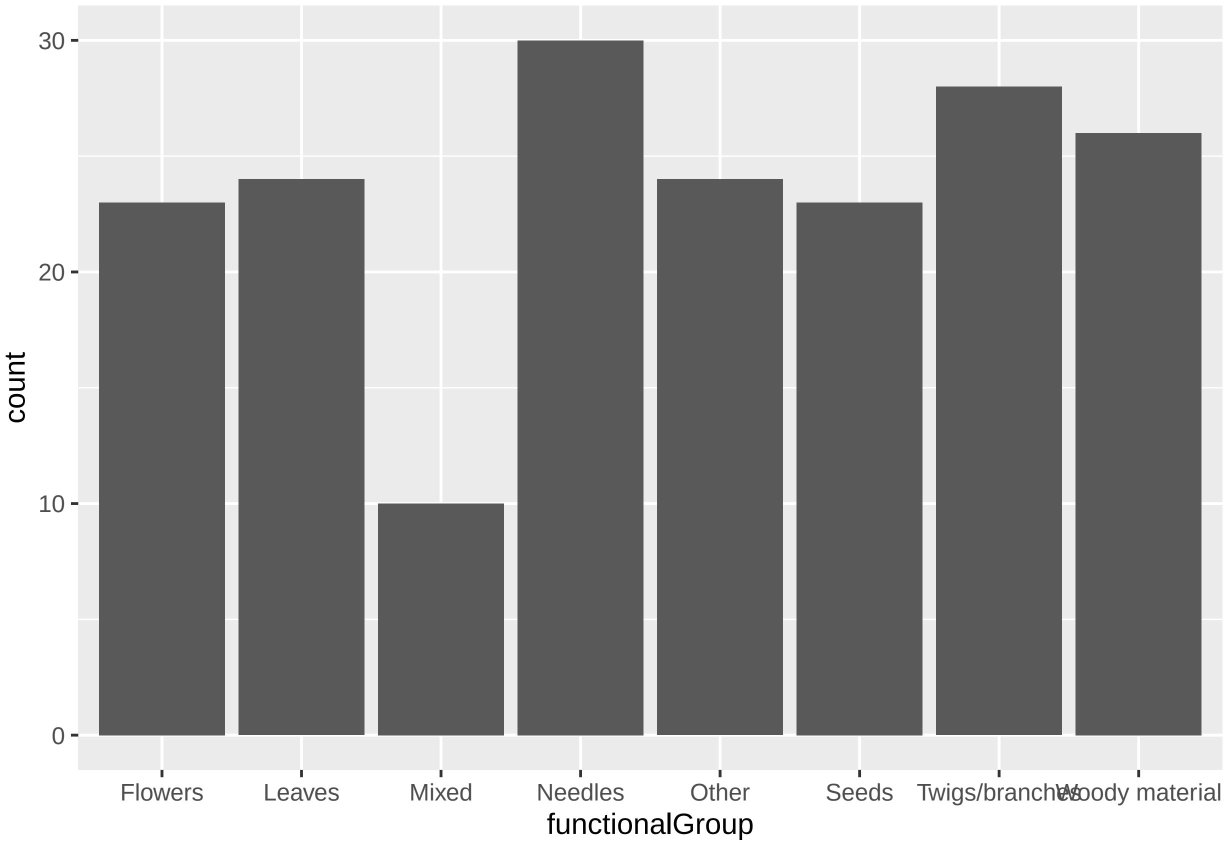
```r
#determine different plots sampled using summary
summary(litter$plotID)
```

```
## 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: The unique function spits out the different plot IDs as a list of levels and then says "12 level." The summary function displays each unique plot ID and how the count for each plot ID.

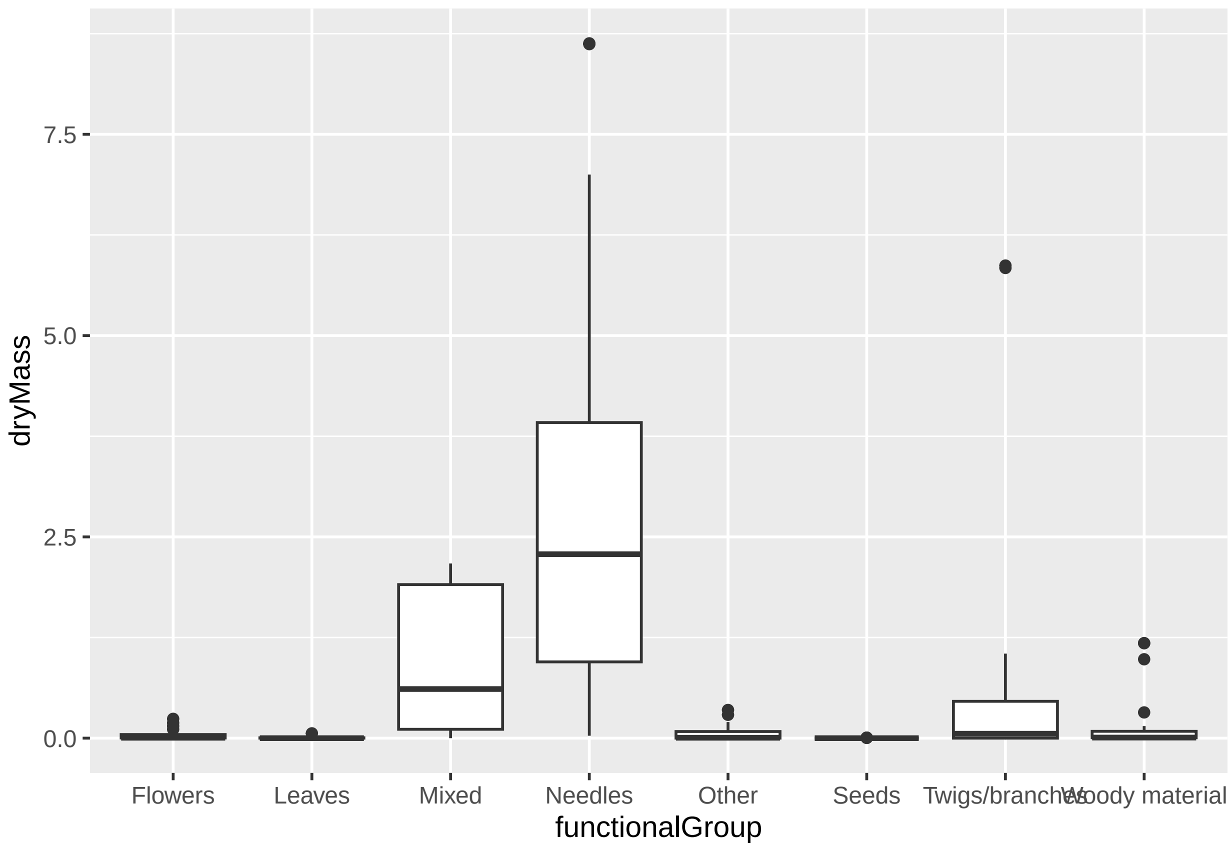
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
#create bar graph of functionalGroup counts
ggplot(data = litter, aes(x = functionalGroup)) +
geom_bar()
```

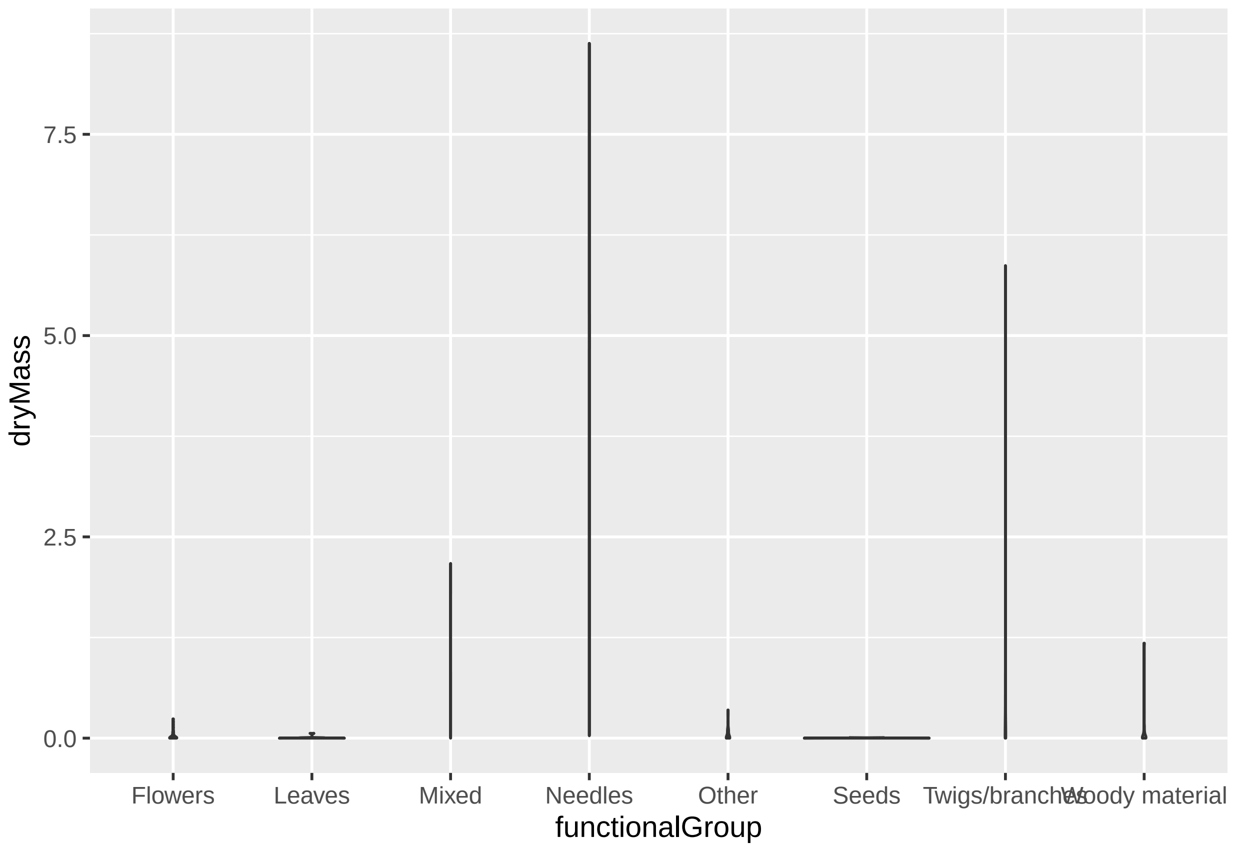


15. Using geom_boxplot and geom_violin, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#create boxplot of dryMass by functionalGroup
ggplot(litter) +
geom_boxplot(aes(x = functionalGroup,y = dryMass,))
```



```
#create violin plot of dryMass by functionalGroup
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: The boxplot is a more effective visualization option than the violin plot because it mainly displays the "center" of the data by displaying the quartiles and mean, and the outliers are not as distracting as they are in the violin plot. There appears to be several outliers, so the violin plot does not do a good job at summarizing the data.

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

> Answer: Mixed, Needeles, and Twigs/branches tend to have the highest biomass at the sites, according to the boxplot.