Assignment 3: Data Exploration

Isabel Zungailia

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.

```
#Check working directory.
getwd()
```

```
## [1] "/home/guest/EDA-Fall2022"
```

```
setwd("/home/guest/EDA-Fall2022")
#install.packages(tidyverse)
library(tidyverse)
#Upload datasets and name them 'Neonics' and 'Litter'.
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: Insecticides have been widely used throughout agriculutre as a means of removing pests or invasive species that might threaten the overall yield or producitivity of the land. It would be beneficial to better understand the ecotoxicology of neonicotinoids on insects to help generate a better understanding of how these insecticides are impacting the insect populations and overall ecosystem health. These understandings can help shape policy reforms moving forward in regard to insecticide use and can help answer questions such as: Are neonicotinoids harmful to the environment? What level of application is considered 'safe' for the insects, plants, and animals in the surrounding ecosystem?

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: Through studying litter and woody debris that falls on the ground in forests, we can gain a better understanding of the fuel load (for forest fire analysis) and further explore the diversity of the forest. This knowledge can better inform bpolicies regarding forest and land management.

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 defined as material that is dropped from the forest canopy and has a butt end diameter of <2cm and a length <50cm, and it is collected in elevated 0.5m^2 PVC traps. Woody debris is material that is dropped from the forest canopy and has a butt end diameter of <2cm and a length >50cm, and it is collected in ground traps (to ensure that larger materials can be collected). 2. One litter trap pair (elevated and ground) is used for every 400m^2 study plot, which totals to around 1-4 traps pairs in each plot. 3. Trap placement within the plots depends on the vegetation present, and can either be random or targeted. Litter trap placement is random when sites have >50% aerial cover of woody vegetation >2m in height. When sites have < 50% cover of woody vegetation, trap placement is targeted to ensure that only areas beneath the qualifying vegetation are considered for placement.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#Observe dimensions of the dataset with 'dim()'.
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?

```
#Use the 'summary' function on the "Effect" column. Determine the most common effects that are studied.
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
```

   Answer: The most common effects that are studied are population, mortality, and behavior. Population and mortality have values over 1000 (population=1803 and mortality=1493), and behavior comes behind them in third with a value of 360. These effects might be of particular interest for this study because they are relatively broad areas that give a good indication as to how neonicotinoids are impacting insects.

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.

```
#Use 'summary' function. Determine the six most commonly used species in the dataset (common name). Com
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 
```

Answer: The six most common species in this dataset include: (1) Honey Bee, (2) Parasitic Wasp, (3) Buff Tailed Honeybee, (4) Carniolan Honey Bee, (5) Bumble Bee, (6) Italian Honeybee. All of these species are types of bees, with the exception of the Parasitic Wasp, and they might be of particular interest in this study over other insects due to their key role in the process of pollination. They contribute to the production of over 1/3 of the world's food supply and are used for various purposes in the field of medicine. Without stable bee populations, we would see major changes in pollination distribution and overall food supply.

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

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

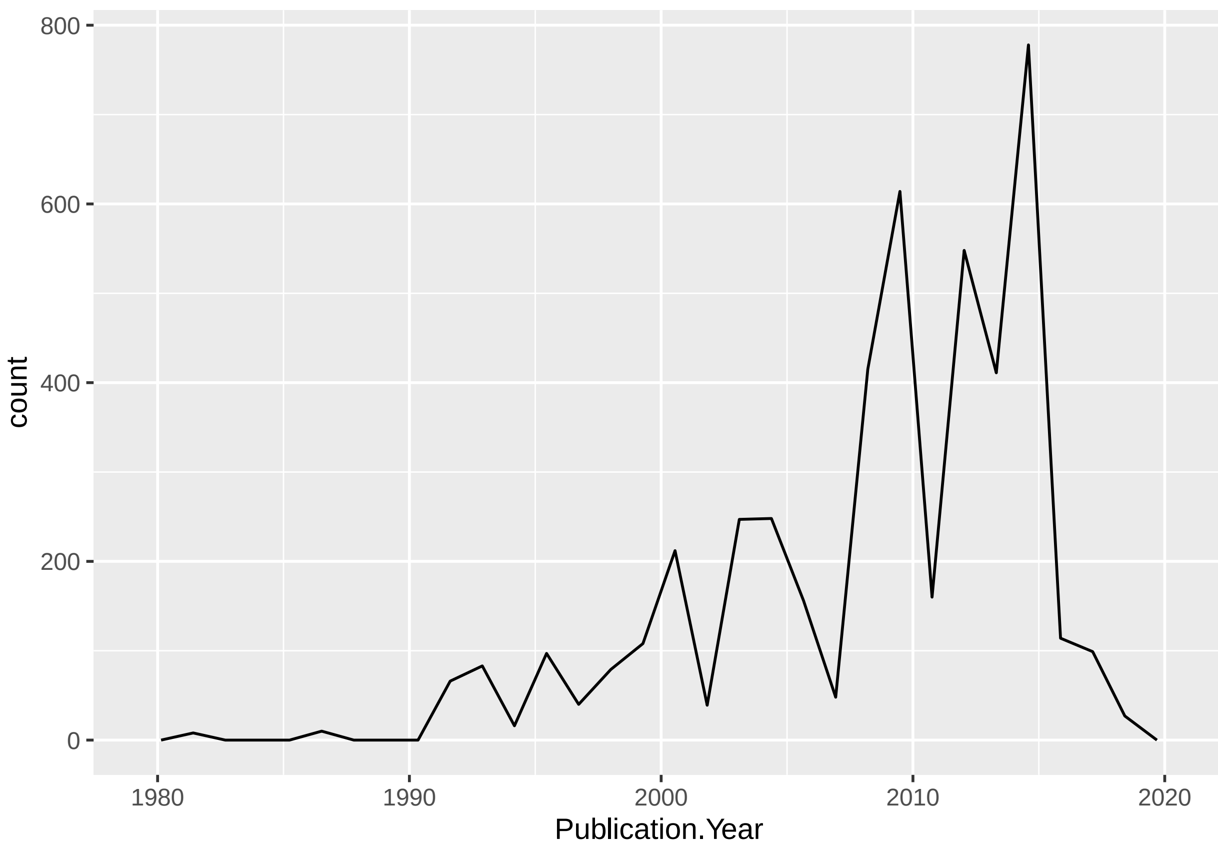
```
## [1] "factor"
```

Answer: The class of 'Conc.1..Author.' in the dataset is a "factor". Factors are variables with discrete values, while numerics are numbers. Even though concentrations should always be a number value, the class of 'Conc.1..Author.' is listed as "factor" because of the 'stringAsFactors' subcommand that we used when loading our datasets, which is used to organize the strings into factors.

Explore your data graphically (Neonics)

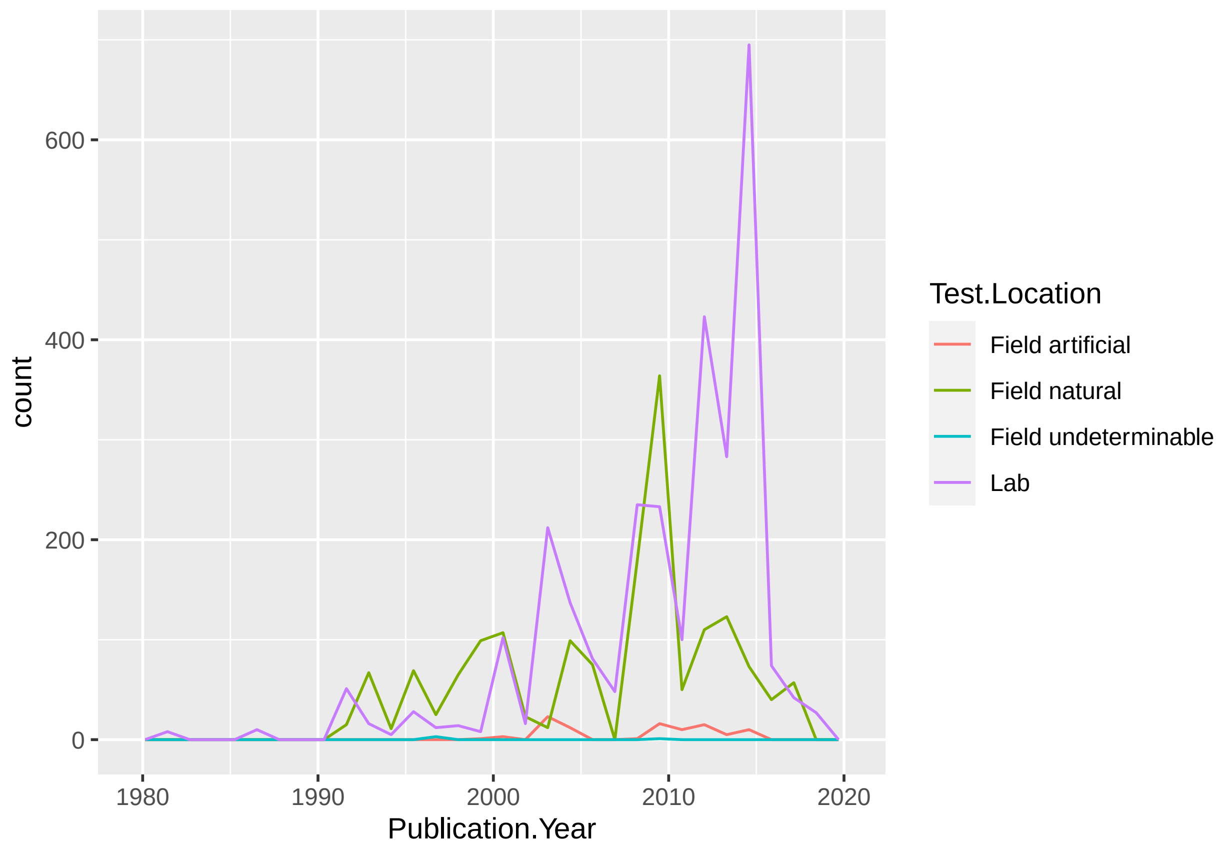
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
#Use function 'geom_freqpoly' to create a plot of studies conducted by publication year.
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 30)
```



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

```
#Recreate the same graph using 'geom_freqpoly'. Change the color aesthetic so that different 'Test.Loca
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 30)
```

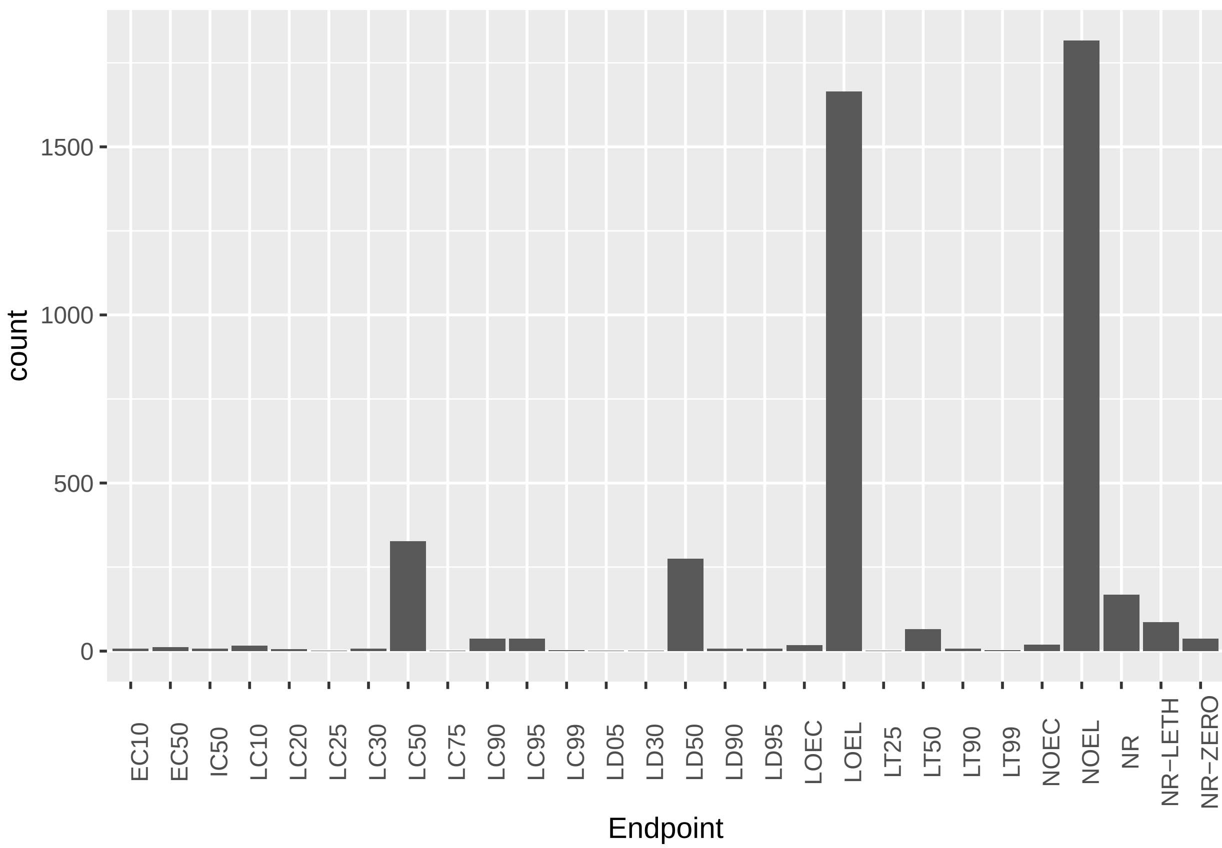


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The two most common test locations are 'Lab' and 'Field natural'. While the test locations vary throughout time, 'Lab' and 'Field natural' are consistently the most commonly used. 'Field articificial' and 'Field undeterminable' have had very low counts throughout the 1980-2020 period. 'Lab' and 'Field natural' both flucatuated between 1990-2020, and 'Lab' had a significant spike in counts just before 2015 (while the counts for the three other test locations remained low during this time).

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.

```
#Create a bar graph using 'Endpoint' counts. Rotate the labels on the x-axis to make them all visible.
ggplot(Neonics, aes(x=Endpoint)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90))
```



Answer: The two most common endpoints are 'NOEL' and 'LOEL'. The 'NOEL' endpoint is part of the "Terrestrial" database and is defined as the "highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test (NOEAL/NOEC)" (pg. 723, Ecotox_CodeAppendix). The 'LOEL' endpoint is also part of the "Terrestrial" database and is defined as the "lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC)" (pg. 722, Ecotox_CodeAppendix). 'NOEL' is the highest-observable-effect-level, while 'LOEL' is the lowest-observable-effect-level.

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.

```
#Determine the class of 'collectDate' variable.
class(Litter$collectDate)
```

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

```
#Change class to a Date.
Litter$collectDate <- as.Date(Litter$collectDate, format= "%Y-%m-%d")
#Confirm new class of the variable.
class(Litter$collectDate)
```

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

```
#Use the 'unique' function to determine which dates litter was sampled in August 2018.
unique(Litter$collectDate)
```

```
## [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`?

```
#Use the 'unique' function to determine how many plots were sampled at Niwot Ridge.
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
```

```
#Use the 'summary' function to compare the outputs.
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
```

```
#Observe the length of 'unique' and 'summary'.
length(unique(Litter$plotID))
```

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

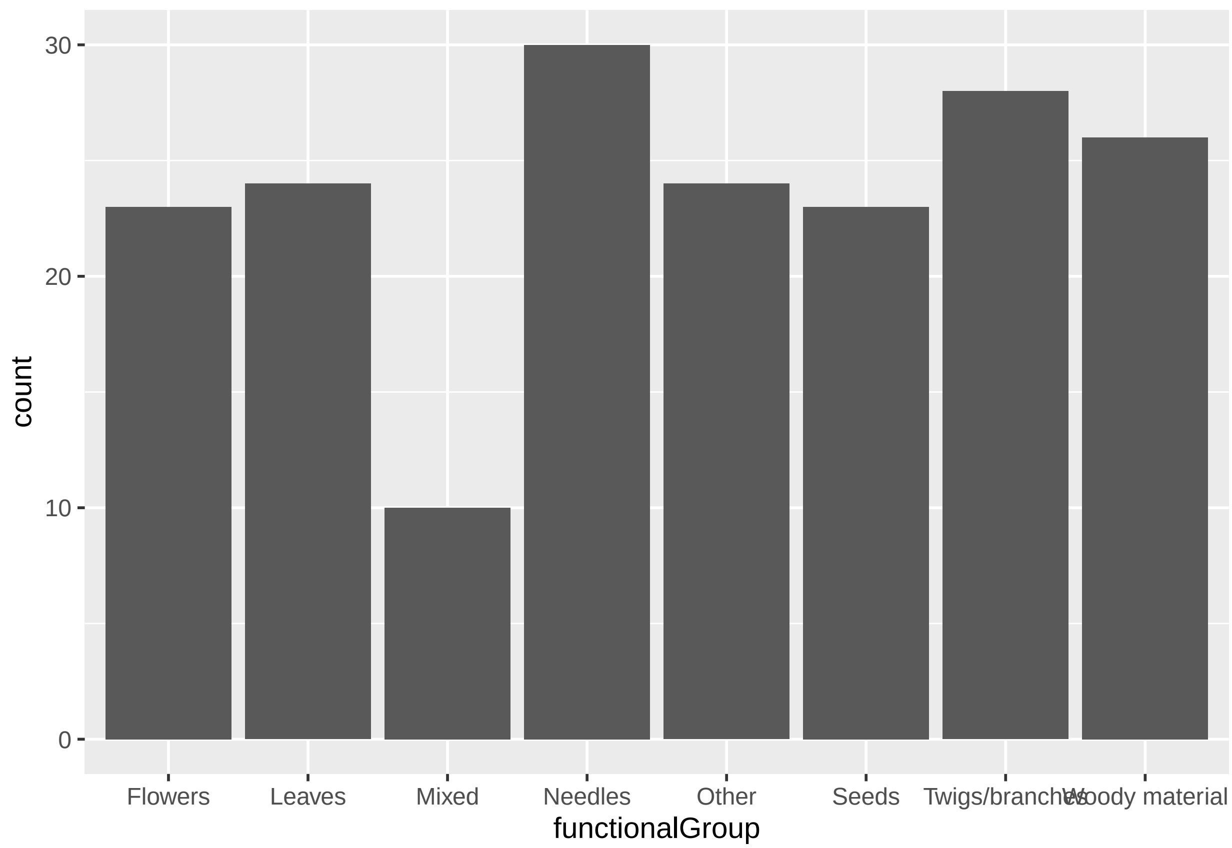
```
length(summary(Litter$plotID))
```

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

> Answer: The 'unique' function presents a list (of plot IDs) of the 12 different plots that were sampled at Niwot Ridge. 'Summary' provides you with a list of the 12 different plots and the number of how many samples were taken from each plot. I used the 'length()' function to observe that both 'unique' and 'summary' gave the same number of outputs (12).

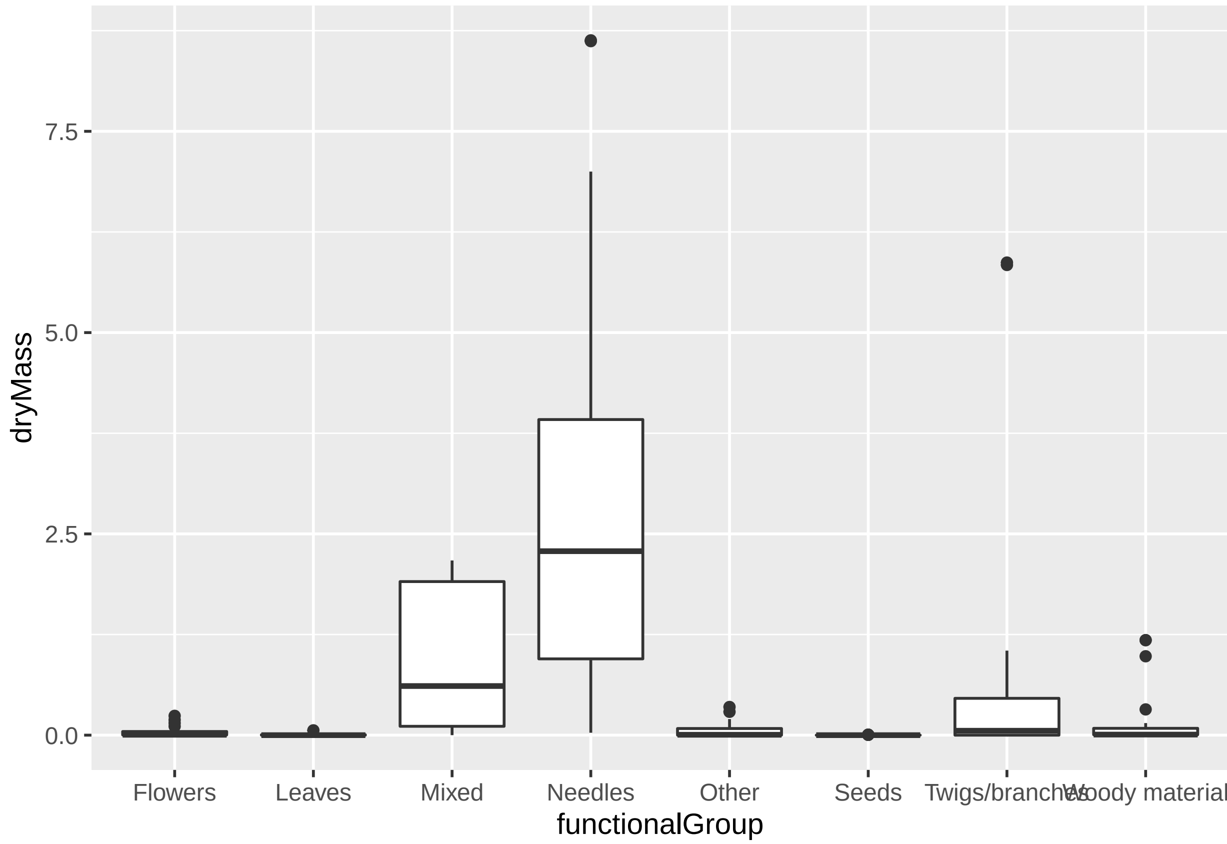
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#Bar graph of 'functionalGroup' counts.
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar()
```

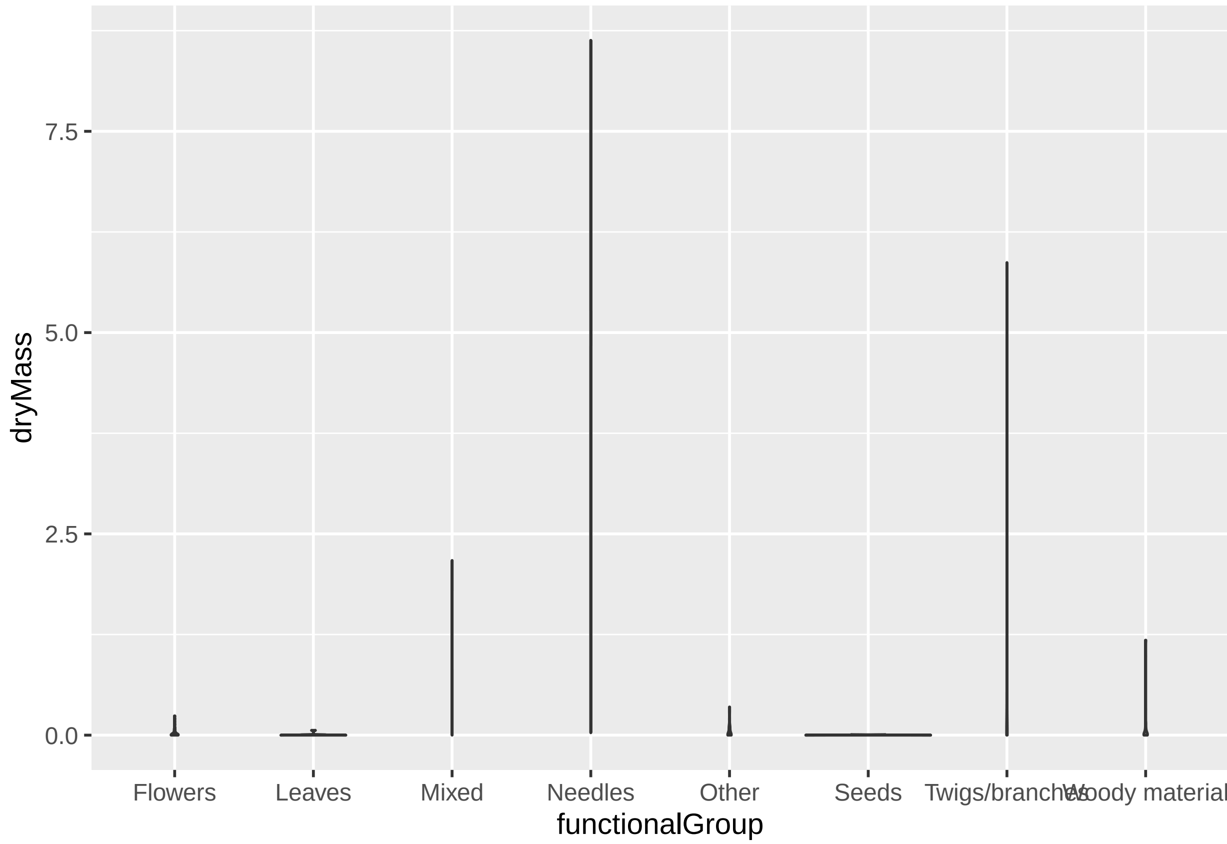


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#Use 'geom_boxplot' to create a boxplot of 'dryMass' by 'functionalGroup'.
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup , y = dryMass))
```




```r
#Use 'geom_violin' to create a 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: In this case, the boxplot provides a more effective visualization of the 'dryMass' by 'functionalGroup'. The boxplot only shows the summary statistics of the dataset (such as mean, median, interquartile range). The violin plot shows the full distribution of the data, so if there is little variance in the dataset, the figure may look odd (which in this case, it does). The boxplot provides the best visualization for the comparison of 'dryMass' by 'functionalGroup' because there is little variance in the dataset.

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

> Answer: At these sites, 'Needles' and 'Mixed' are the two litter types with the greatest biomass at these sites.