Assignment 3: Data Exploration

Ashton Cloer

Fall 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.

```
library(tidyverse) #loading neccesary packages
library(lubridate)
library(here)
library(ggplot2)

getwd() #setting working directory
```

```
## [1] "/Users/ashtoncloer/EDE_Fall2023"
```

```
here()
```

```
## [1] "/Users/ashtoncloer/EDE_Fall2023"
```

```
Neonics.df <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = TRUE)
Litter.df <- 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: Depending on how neonicotinoids effect certain insect populations, application of the chemical in agriculture could have negative impacts on ecosystem food webs. Neonicotinoids could kill certain insect species that are important food sources for protected bird species or the chemical could bioaccumulate in birds and other animals that eat insects. Or, the neonicotinoids could kill valuable insects like pollinators.

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: From the standpoint of carbon offsets, measuring the amount of leaf litter and woody debris is a potential method for calculating the amount of carbon sequestration a certain forest ecosystem can achieve. Measuring leaf litter and woody debris can also provide insights to the nutrient cycles and levels of carious ecosystems.

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.In sites with forested tower airsheds, itter sampling takes place in 20 40m x 40m plots. 2. In sites with low-saturated vegetation, litter sampling takes place in 4 40m x 40m plots and 26 20m x 20m plots. 3. Trap placement within plots is either targeted or randomized. Sites with more than 50% aerial cover of woody vegetation taller than 2m are randomized. All other sites (those with less than 50% cover of woody vegetations and/or sites that are made of heterogeneous vegetation) are targeted to account for specific vegetation.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics.df)
```

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

Answer: 4632 observations of 30 columns/variables

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?

```
summary(Neonics.df$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: Mortality and population are the most commonly studied effects. Given neonicotinoid is an insecticide, it makes sense that the researchers would collect the most information on population and mortality to understand how the chemical impacts mortality rates amongst various insect populations.

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...]

```
most_common <- summary(Neonics.df$Species.Common.Name)

sort(most_common) #sorting the summary output
```

```
##                          Ant Family                         Apple Maggot
##                                   9                                    9
##              Glasshouse Potato Wasp                             Lacewing
##                                  10                                   10
##             Southern House Mosquito             Two Spotted Lady Beetle
##                                  10                                   10
##           Spotless Ladybird Beetle                 Braconid Parasitoid
##                                  11                                   12
##                        Common Thrip        Eastern Subterranean Termite
##                                  12                                   12
##                              Jassid                           Mite Order
##                                  12                                   12
##                           Pea Aphid                   Pond Wolf Spider
##                                  12                                   12
##               Armoured Scale Family                   Diamondback Moth
##                                  13                                   13
##                       Eulophid Wasp                  Monarch Butterfly
```

```
##                              13                                13
##                   Predatory Bug              Yellow Fever Mosquito
##                              13                                13
##                    Corn Earworm                  Green Peach Aphid
##                              14                                14
##                       House Fly                         Ox Beetle
##                              14                                14
##              Red Scale Parasite                 Spined Soldier Bug
##                              14                                14
##           Western Flower Thrips Hemlock Woolly Adelgid Lady Beetle
##                              15                                16
##           Hemlock Wooly Adelgid                              Mite
##                              16                                16
##                     Onion Thrip              Araneoid Spider Order
##                              16                                17
##                       Bee Order                    Egg Parasitoid
##                              17                                17
##                    Insect Class           Moth And Butterfly Order
##                              17                                17
##     Oystershell Scale Parasitoid          Black-spotted Lady Beetle
##                              17                                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
##                    Codling Moth         Flatheaded Appletree Borer
##                              19                                20
##             Horned Oak Gall Wasp                 Leaf Beetle Family
##                              20                                20
##                Potato Leafhopper         Tooth-necked Fungus Beetle
##                              20                                20
##                   Argentine Ant                            Beetle
##                              21                                21
##                      Mason Bee                          Mosquito
##                              22                                22
##                Citrus Leafminer                   Ladybird Beetle
##                              23                                23
##               Spider/Mite Class                Tobacco Flea Beetle
##                              24                                24
##                    Chalcid Wasp             Convergent Lady Beetle
##                              25                                25
##                   Stingless Bee               Ground Beetle Family
##                              25                                27
##               Rove Beetle Family                      Tobacco Aphid
##                              27                                27
##                   Scarab Beetle                     Spring Tiphia
##                              29                                29
##                     Thrip Order             Ladybird Beetle Family
##                              29                                30
##                      Parasitoid                     Braconid Wasp
```

```
##                              30                                33
##                     Cotton Aphid                   Predatory Mite
##                              33                                33
##               Sweetpotato Whitefly                     Aphid Family
##                              37                                38
##                    Cabbage Looper             Buff-tailed Bumblebee
##                              38                                39
##                    True Bug Order           Sevenspotted Lady Beetle
##                              45                                46
##                      Beetle Order       Snout Beetle Family, Weevil
##                              47                                47
##                Erythrina Gall Wasp                Parasitoid Wasp
##                              49                                51
##             Colorado Potato Beetle                   Parastic Wasp
##                              57                                58
##               Asian Citrus Psyllid                Minute Pirate Bug
##                              60                                62
##                  European Dark Bee                         Wireworm
##                              66                                69
##                   Euonymus Scale                Asian Lady Beetle
##                              75                                76
##                  Japanese Beetle                 Italian Honeybee
##                              94                               113
##                       Bumble Bee              Carniolan Honey Bee
##                             140                               152
##              Buff Tailed Bumblebee                  Parasitic Wasp
##                             183                               285
##                        Honey Bee                         (Other)
##                             667                               670
```

Answer: The honey Bee, parasitic wasp, buff-tailed bumblebee, carniolan honey bee, bumble bee, and italian honeybee are the most commonly studied species, despite a category that reference 'other' species and totals 670 observations. The 5 bee species listed above are all pollinators and are important for humans and ecosystems alike. Studying the impact of neonicotinoid on pollinators is important for ensuring continued cross-pollination and subsequent growth of various plants used in agriculture. The parasitic wasp is also important for agriculture as they are natural methods for controlling other pests that often eat/ruin crops.

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.df$Conc.1..Author.)
```

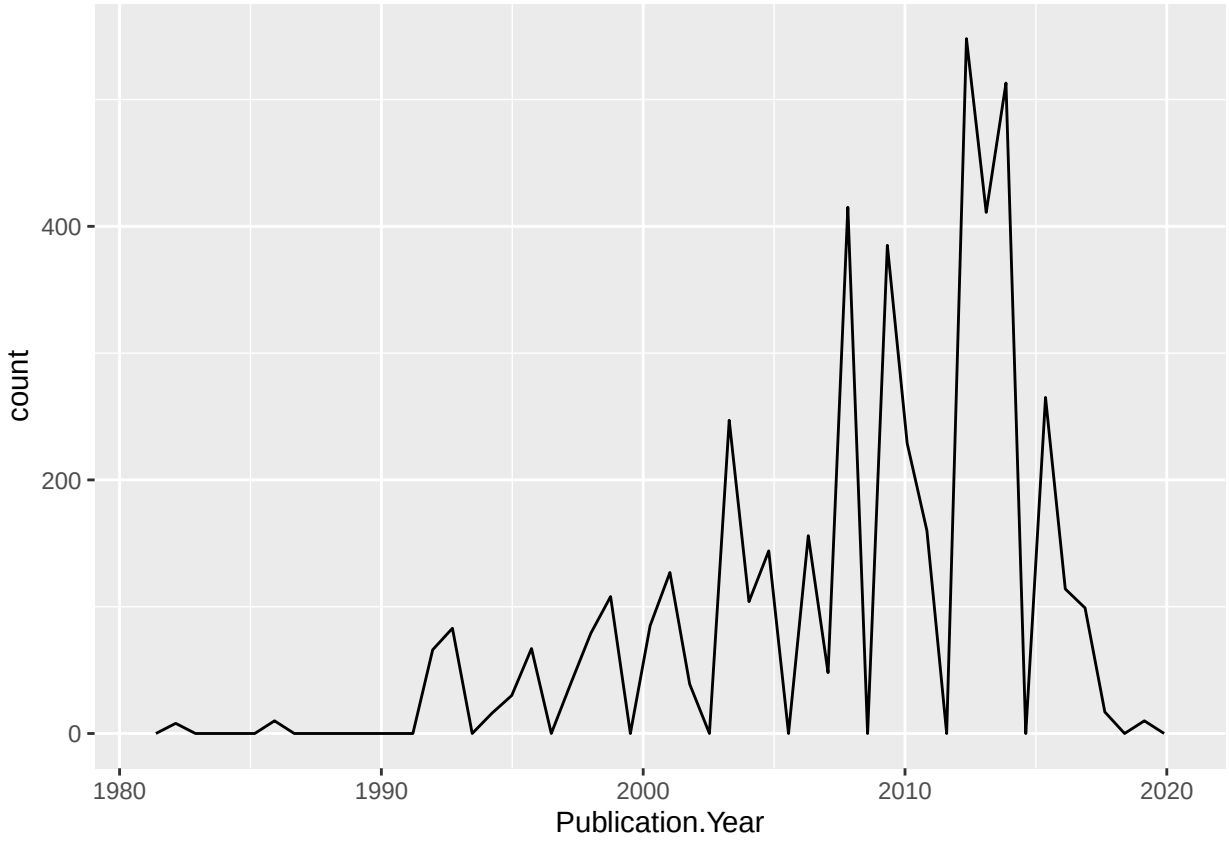
```
## [1] "factor"
```

Answer: It is a factor class. I am not sure why R is reading it as a factor class rather than a character class, perhaps because the majority of responses are simply numbers. But the column also includes results such as numbers with / at the end to represent 'per' unit and 'NR' for non-reported values instead of NA making it a factor class rather than numeric.

Explore your data graphically (Neonics)

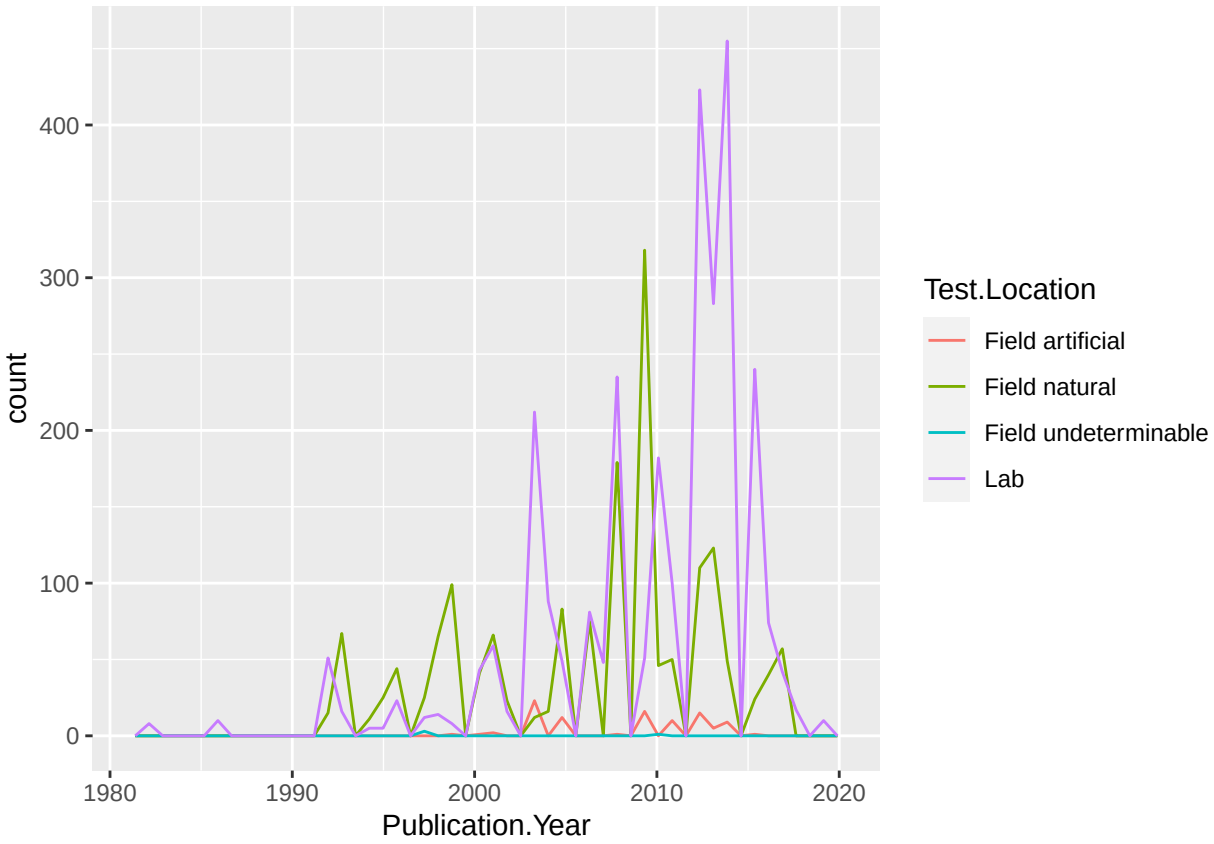
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics.df) +
  geom_freqpoly(aes(x = Publication.Year), bins=50) #setting x-axis and bins
```



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

```
ggplot(Neonics.df) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins=50)
```



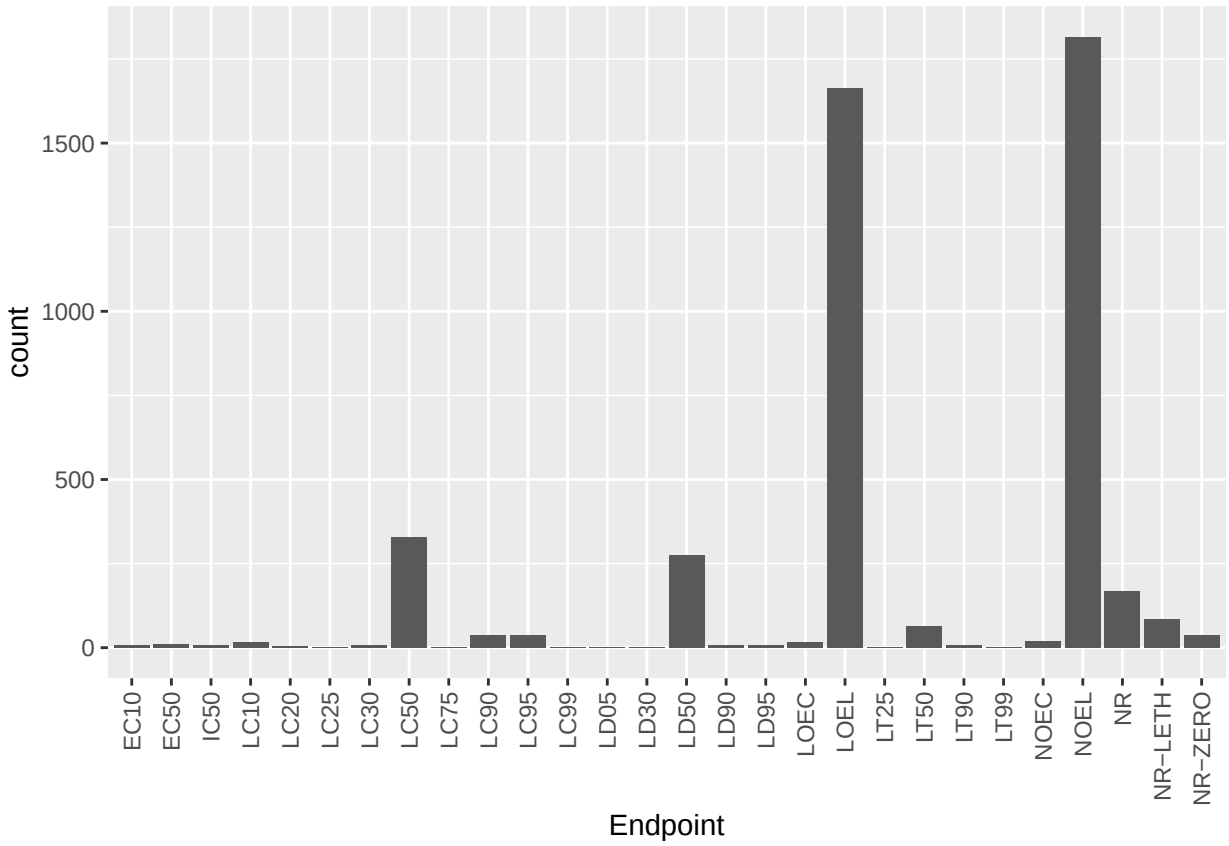
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are the lab and the natural field. At first (early 90s through around 2005), occurences were relatively equal in both the lab and the natural field. But around 2010, the field natural observations were much more frequent. A few years later, around 2012, the lab observations began heavily outweighing any other mode of observations.

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.df, aes(x = Endpoint)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) #rotating and aligning x-axis l
```



Answer: The two most common endpoints are LOEL and NOEL. According to the Ecotox appendix, LOEL refers to the lowest-observable-effect-level or the lowest dose that produced effects that were signficantly different from the responses resulting from controls. NOEL refers to no-observable-effect-level or the highest dose producing effects not signficantly different from the responses resulting from 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.

```
class(Litter.df$collectDate) #factor
```

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

```
Litter.df$collectDate <- as.Date(Litter.df$collectDate, format = "%Y-%m-%d")

class(Litter.df$collectDate) #date
```

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

```
unique(Litter.df$collectDate) #sampled in August 2018 on the 2nd and 30th
```

```
## [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.df$plotID) #12 different plots
```

```
##  [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
```

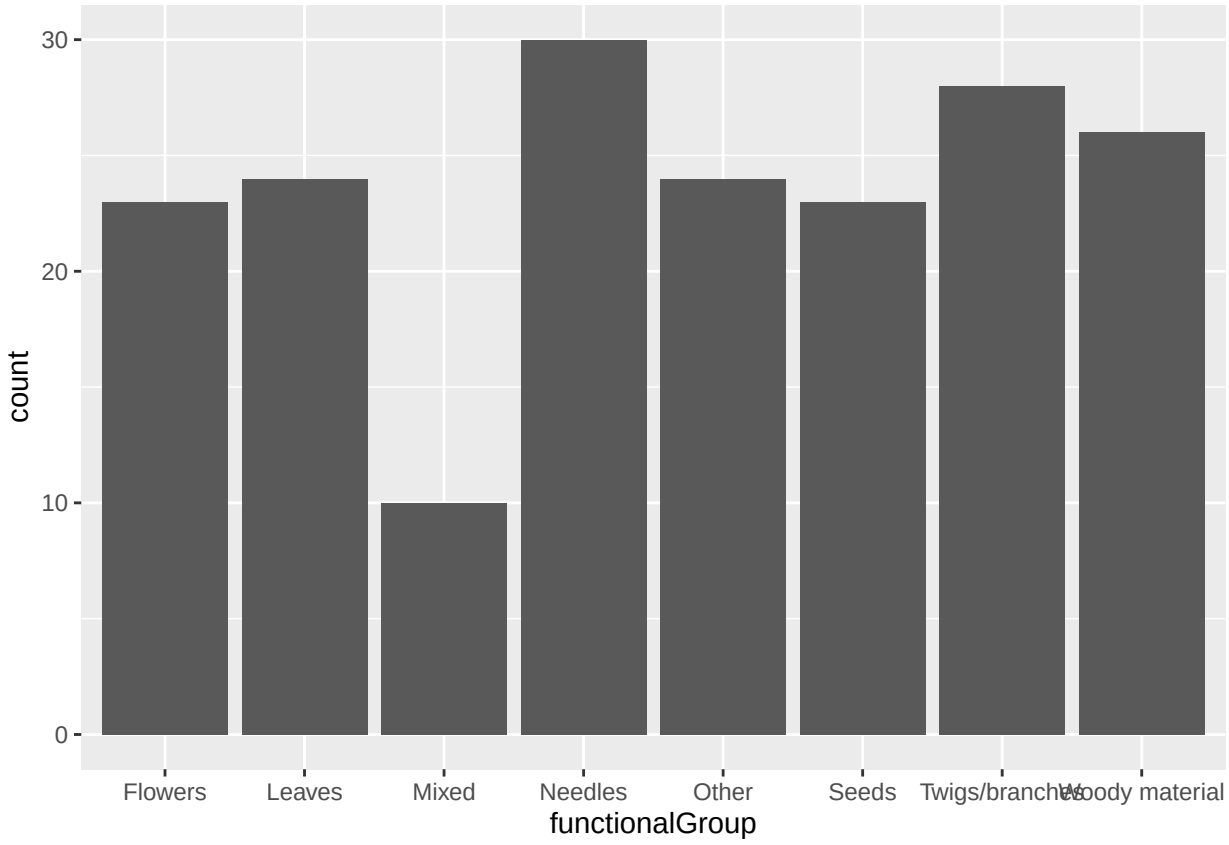
```
summary(Litter.df$plotID)
```

```
## 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. The function summary gives a county of each of the different plot numbers where as the unique function only shows the various types of plot numbers without a count.

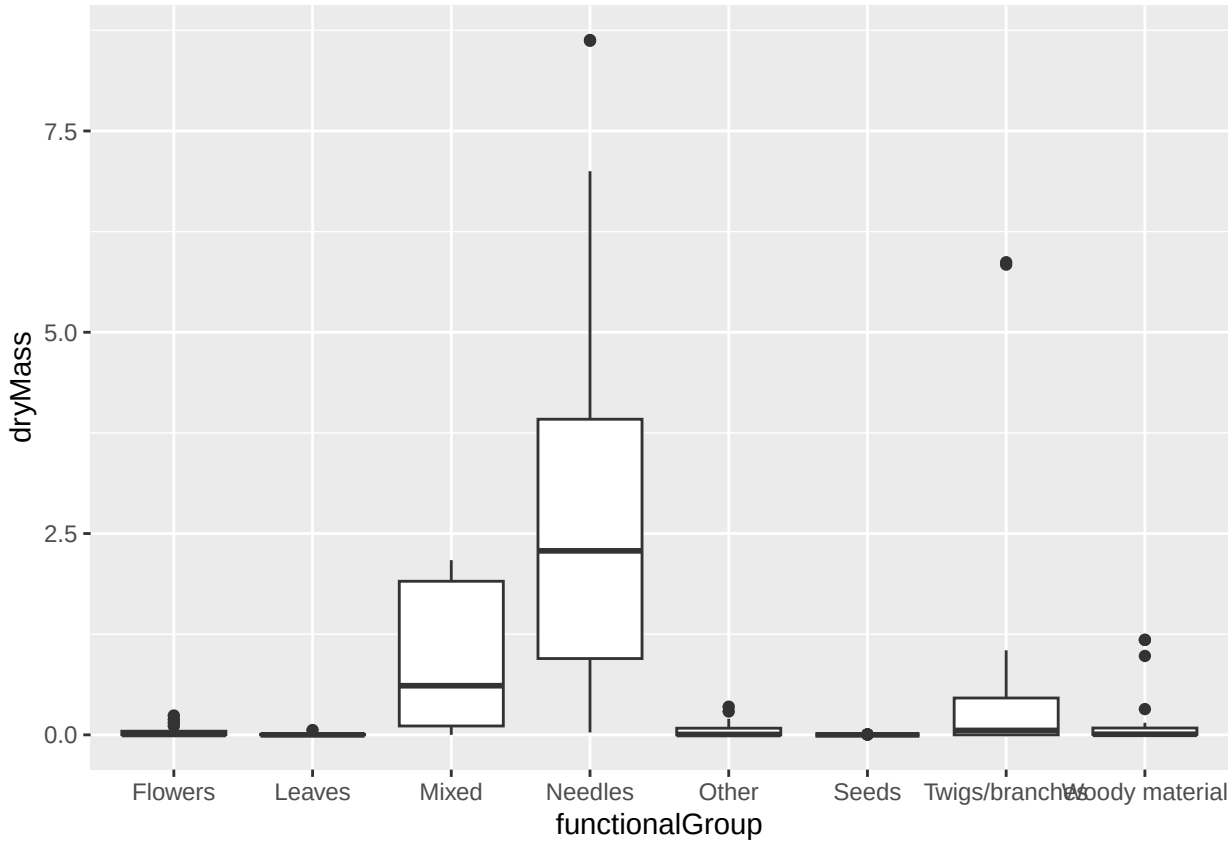
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter.df, aes(x = functionalGroup)) +
  geom_bar()
```

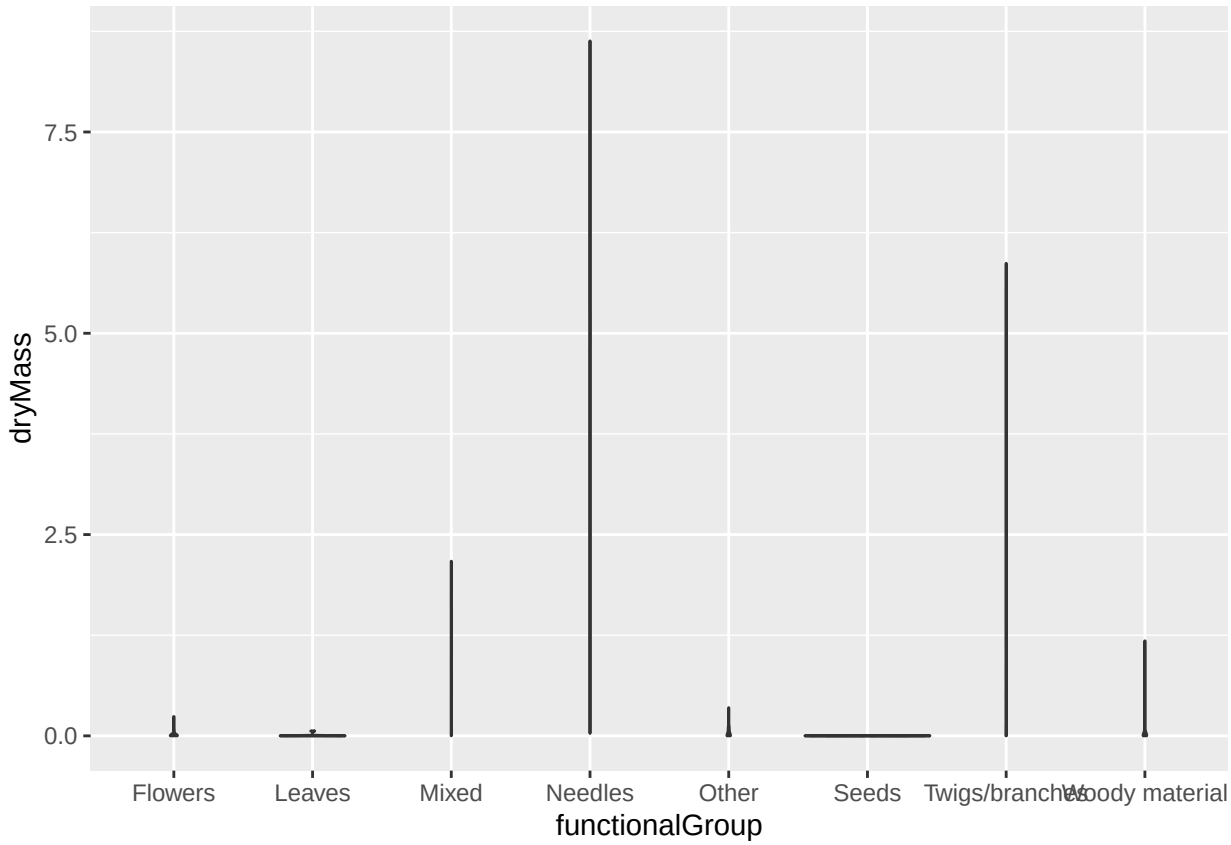


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(Litter.df) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass))
```




```
ggplot(Litter.df) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75))
```



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

> Answer: The boxplot shows more details regarding the summary statistics of the various functional groups as opposed to the violin plot, which seems to only show information on the distribution and density of dry mass values by functional groups.

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

> Answer: According to the box-and-whisker plot, needles and mixed litter tend to have the highest interquartile range of biomass, followed by twigs and branches. However, the twigs/branches group has an outlier that is larger than the mixed litter, which shows that needles, twigs and branches have the highest biomass. Opposingly, the violin chart shows the overall range of dry mass recorded for each functional group and shows needles and twigs/branches to have the overall highest biomass.