Assignment 3: Data Exploration

Matthew Vining

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.

```
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=80), tidy=TRUE)
getwd()
```

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

```
#install.packages(tidyverse)
library(tidyverse)
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 are applied to many crops in agriculture to help increase their crop yield and health by protecting them from pests such as insects that can damage or consume the crops if not properly protected. We may be interested in the ecotoxicology of this substance on insects to determine if it can effectively kill and remove these pests, as well as understanding the relationship

between its effect on killing insects while also ensuring it doesn't contaminate other parts of the environment/kill valuable insects we do not want to damage.

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:The use of forest litter and woody debris, once decomposed, is a good source of nutrients and organic matter for species growing in the forest. We may be interested in it for several reasons. Researchers may be interested in the health of these forest biomes based on the amount, mass, and quality of the litter and debris in the areas. They may also be interested in how it affects soil quality. They may be even interested in its affect on creating habitats for certain organisms.

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: It is sampled in a particular way to yield desired results. A few of these method details are: 1. Litter and woody debris are collected from elevated and ground traps, respectively. 2.Mass data for samples are measured separately for several functional groups (leaves, needles, etc), to an accuracy of 0.01 grams. 3. Litter and woody debris sampling is done at terrestrial NEON sites that contain woody vegetation greater than 2m tall.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
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?

```
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 studied are mortality and population. They are probably of specific interest because we can see the number of insects/insect species sampled and the resulting mortality rate of them once a neonicotinoid is applied to a crop. Before and after effects of treatment.

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.

```
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 are (not including other as a category) honey bee, parasitic wasp, buff tailed bumblebee, carniolan honey bee, bumble bee, and italian honeybee. All of the most common specieis are species of bees. It is likely they are the most researched/studied because they are vulnerable species of insect that are needed to polliinate many plants, including agricultural crops. If neonicotinoids are meant to remove and kill insects, and bees are important to ecosystem funcitons, we care about studying them more to understand how to protect them from certain types of neonicotinoids.

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

```
class(Neonics$Conc.1..Author.)
```

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

```
summary(Neonics$Conc.1..Author.)
```

```
##    0.37/      10/      NR/       NR        1     1023    0.40/       2/
##      208      127      108       94       82       80       69       63
##       10   0.053/      100      50/     0.5/     0.03    0.05/     0.45
##       62       59       56       51       45       44       43       43
##      0.1/    0.45/     1.0/    2.27/       50    0.125     500/      0.5
##       42       40       40       40       36       33       33       32
##   0.048/    0.15/       1/       48    25.0/      12/    0.027      2.4
##       30       30       30       30       28       27       26       26
##      0.2/    0.56/     100/        3    0.01/    1000/       3/    0.336
##       25       24       23       23       22       22       22       21
##      1.5/     0.05      1.5    2.60/    20.0/        6    6.80/    62.5/
##       21       20       20       20       20       20       20       20
##    0.005     0.4/    0.18/     0.3/     1000       40 0.00355/      0.1
##       18       18       17       17       17       17       16       16
##      0.4     150/      300      80/    0.053     0.24     0.28     125/
##       16       16       16       16       15       15       15       15
##        9   0.0001  0.0004/   0.084/     0.15      0.6    12.5/   144.0/
##       15       14       14       14       14       14       14       14
##     350/    40.0/      48/       56      84/    0.17/      125       14
##       14       14       14       14       14       13       13       13
##       16       17   0.047/    0.25/    0.28/    1.28/    1.81/      112
##       13       13       12       12       12       12       12       12
##      150     2.5/       25      60/      75/    0.02/   0.025/     0.29
##       12       12       12       12       12       11       11       11
##    37.5/       4/        5  (Other)
##       11       11       11     1817
```

Answer: The class of Conc.1..Author is a factor. Based off the naming structure and summary of the variable, it appears as a factor maybe because its a concentration range type that is assigned to different observations/individual measurements, rather than an actual measure of concentration that varies more widely. This would make it easier to group observations together depending on their concentration type.

Explore your data graphically (Neonics)

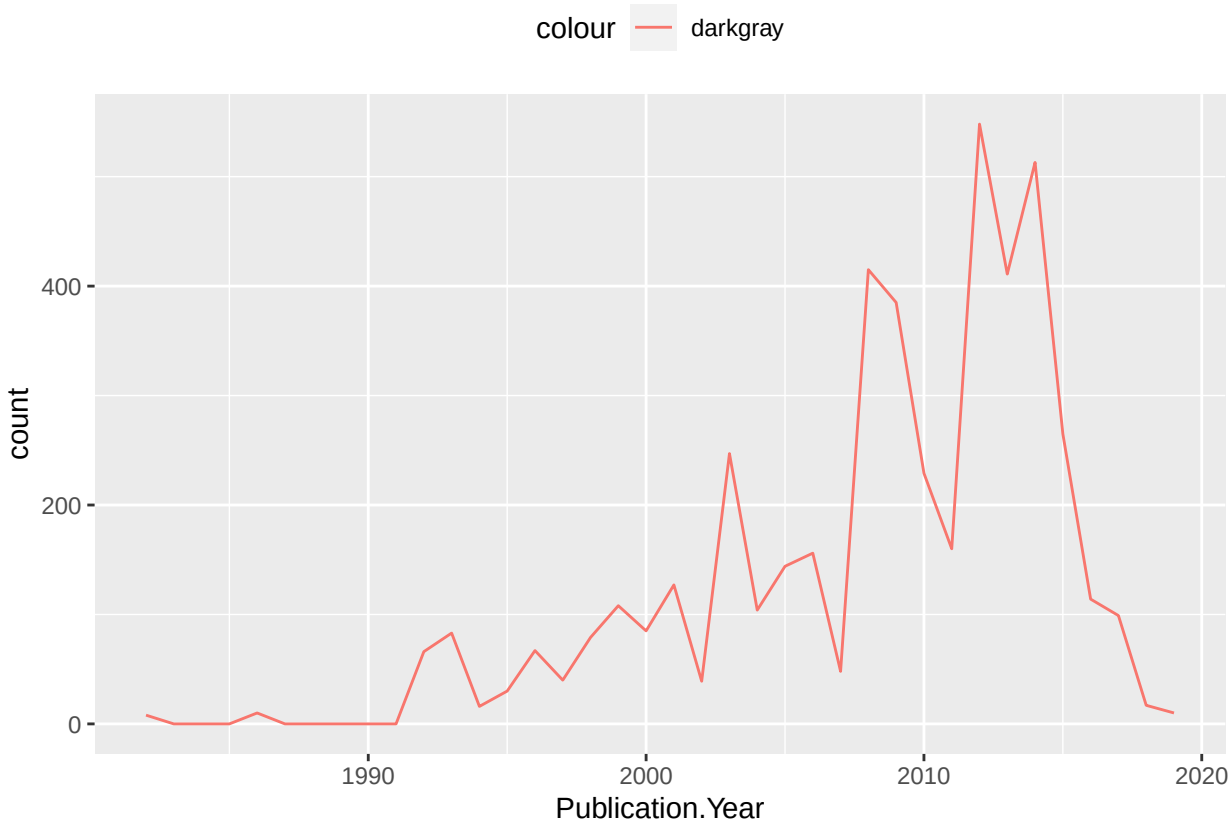
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
summary(Neonics$Publication.Year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1982    2005    2010    2008    2013    2019
```

```
ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year, color = "darkgray"), bins = 38) +
    scale_x_continuous(limits = c(1982, 2019)) + theme(legend.position = "top")
```

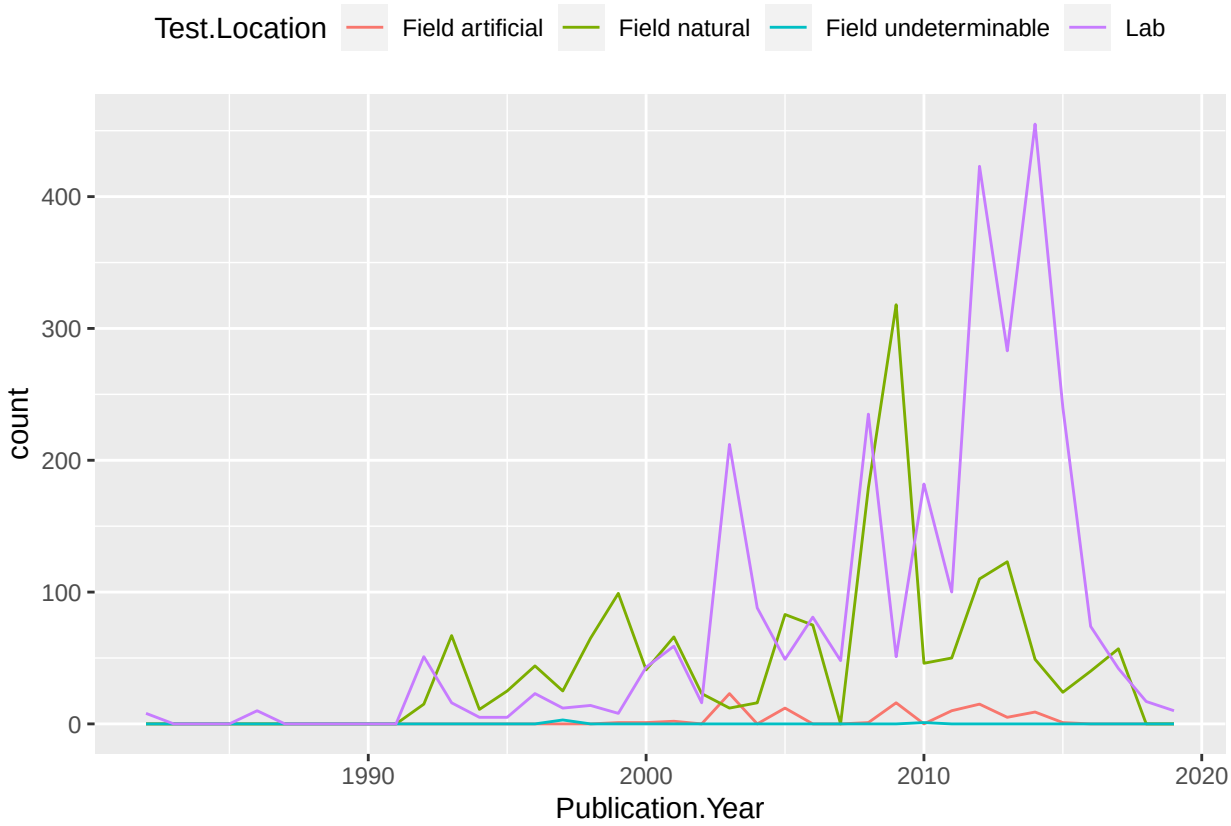
```
## Warning: Removed 2 row(s) containing missing values (geom_path).
```



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),
    bins = 38) + scale_x_continuous(limits = c(1982, 2019)) + theme(legend.position = "top")
```

```
## Warning: Removed 8 row(s) containing missing values (geom_path).
```

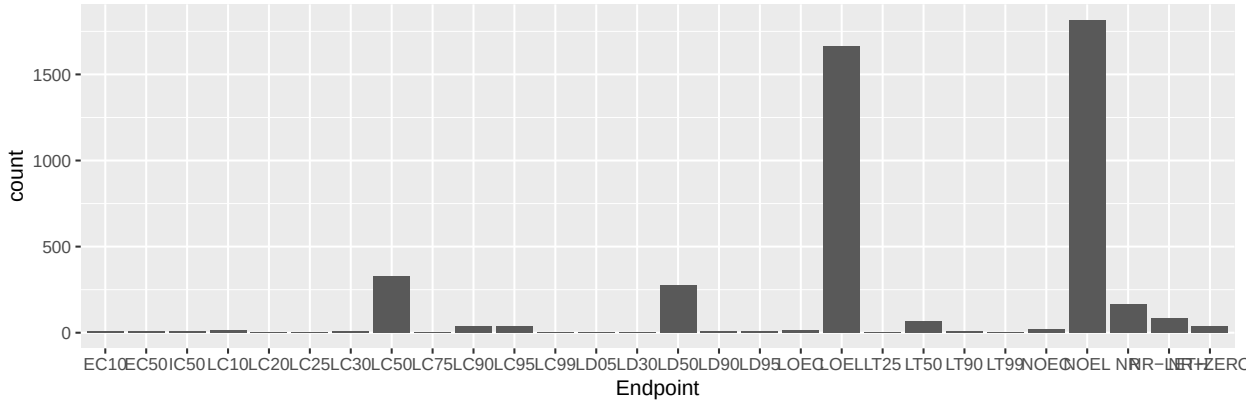


Interpret this graph. What are the most common test locations, and do they differ over time?

Answer:Overall it appears that the most common test location is in the lab over time. After that, the most common is in the natural field setting, with a high spike above lab settings in the 2007-2008 time period. As time has progressed for this sample, lab settings have become the most common.

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.

```
ggplot(Neonics, aes(x = Endpoint)) + geom_bar()
```



Answer:The two most common endpoints are LOEL and NOEL. LOEL is defined as "Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly

different (as reported by authors) from responses of controls (LOEAL/LOEC)". 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)".

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)
```

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

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

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

```
unique(Litter$collectDate)
```

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

```
# litter sampled on the 2nd and 30th of August, 2018.
```

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
```

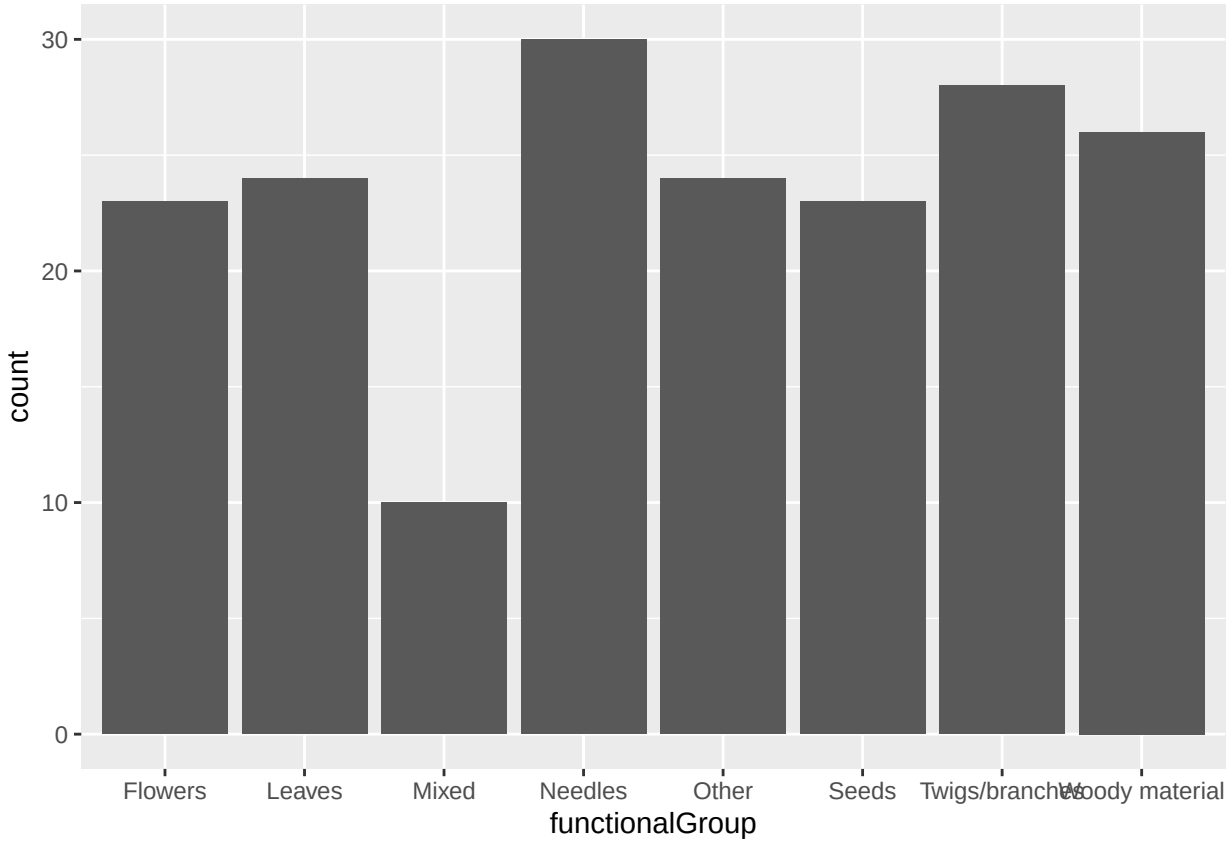
```
summary(Litter$plotID)
```

```
## 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: Using unique, we can see that 12 different plots were sampled. The difference between unique and summary is that unique pulls the number of unique values in a given variable, and removes any duplicates. Summary shows the amount of duplicated values per variable. So if we wanted to know the amount of times each plot was sampled, then summary would be a better function to use. If we simply want to know the number of plots that were sampled, then the unique function is better.

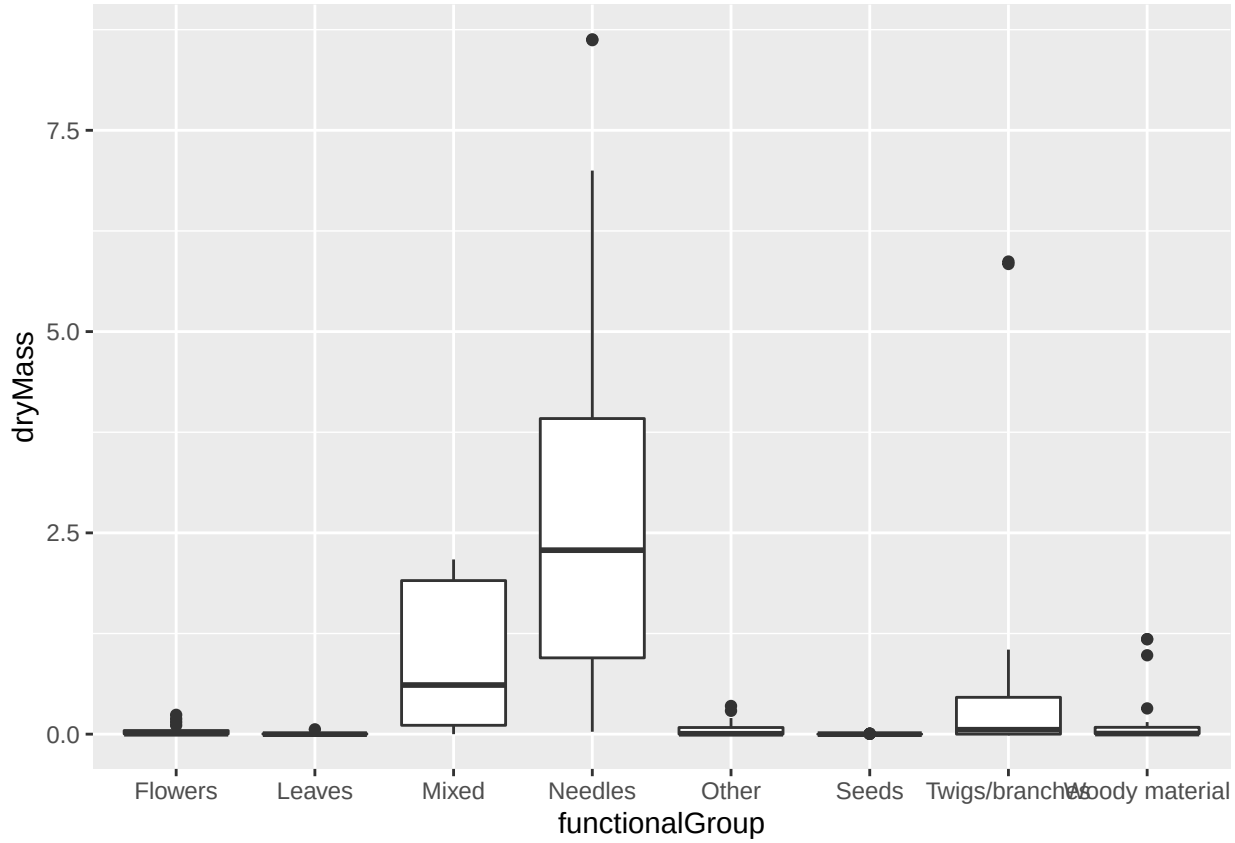
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

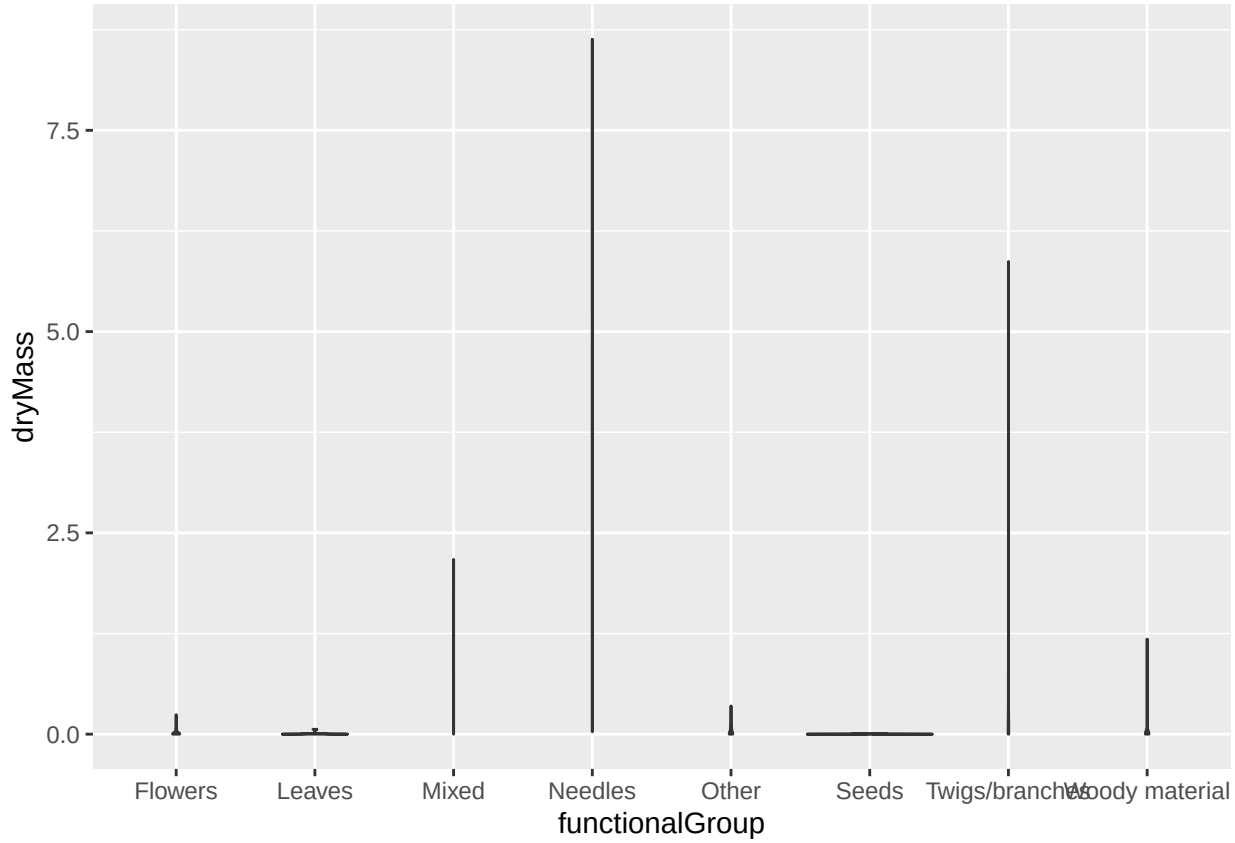


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




```
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: Because the litter types are fairly equally distributed across the Niwot Ridge sites, there is not much variation, something a violin plot highlights. A violin plot performs the same as a boxplot but adds in a factor of density distributions; however, becuase the sampling count per functional group was pretty even, it doesn't show well.

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

> Answer:Looking at both the median and maximum values, needles tend to have the highest biomass at these sites. After that mixed materials, and then twigs and branches.