Assignment 3: Data Exploration

Elizabeth Good

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.

```
getwd() # check working directory
```

```
## [1] "C:/Users/goode/OneDrive/Documents/Duke/ENV872_EDE/EDE_Fall2023"
```

```
library(tidyverse) # load in necessary packages
library(lubridate)

Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",stringsAsFactors = TRUE)
# upload the neonic dataset

Litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv",stringsAsFactors = TRUE)
# upload the litter and woody debris dataset
```

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: Since neonicotinoids are insecticides, which means a chemical that has the purpose of killing insects, it is important to know how effective the neonicotinoids are at their job. Ecotox studies will provide information like what kind of dose and what duration of exposure/application methods are most effective at killing the desired species of insect, which is important knowledge for farmers or others trying to use the neonicotinoids in their agricultural work.

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: After a quick search of some publications from the USFS, I learned that woody debris in forests is important due to its role in the carbon budges and nutrient cycling of forests. Woody debris also provides a source of energy for aquatic ecosystesm and habitat for both terrestrial and aquatic organisms. It can also influence water flow and sediment transport.

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. Sampling take place in terrestrial sites where there is woody vegetation that is more than 2m tall.
   2. Depending on vegetation coverage, the placement of traps can be random (sites with greater than 50% woody vegetation coverage) or targeted (sites with less than 50% woody vegetation coverage). 3. In terms of temporal sampling, ground traps get samples once per year while elevated traps can get sampled once every 2 weeks or once every 1-2 months depending on the forest type.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
dim(Neonics)
```

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

```r
# finding the dimensions of the Neonics data set; 4623 rows and 30 columns
```

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?

```r
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 
```

```r
# finding how many of each type of effect are in the data set
```

> Answer: Population is the most common effect studied (1803 counts) followed closely by mortality (1493 counts). Mortality refers to endpoints where the cause of death was directly related to the chemical of interest, and population refers to endpoints relating to a group of the same species in the same place at the same time. These are very important effects to study in insects because it will show if the chemical will perform its job and kill the insect and how the chemical effects a group of a particular species of 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.[TIP: The `sort()` command can sort the output of the summary command…]

```r
sort(summary(Neonics$Species.Common.Name))
```

```
##                         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
```

```
# finding how many of each type of species are in the data set and
# ordering them from lowest to highest count
```

Answer: Not including “other” the six most commonly studeid species are honey bees, parasitic wasps, buff tailed bumblebees, Carniolan honey bees, bumble bees, and Italian Honeybees. These are all pollinators and are crucial for healthy ecosystems and successful agricultural opperations (wasps are pollinators but are also beneficial becuase they can kill pests that would be harmful to crops). None of these species are harmful to crops and are beneficial to crops and ecosystems. There are likely lots of studies of these species to ensure that the insecticides aren’t harmful to or target these species.

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

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

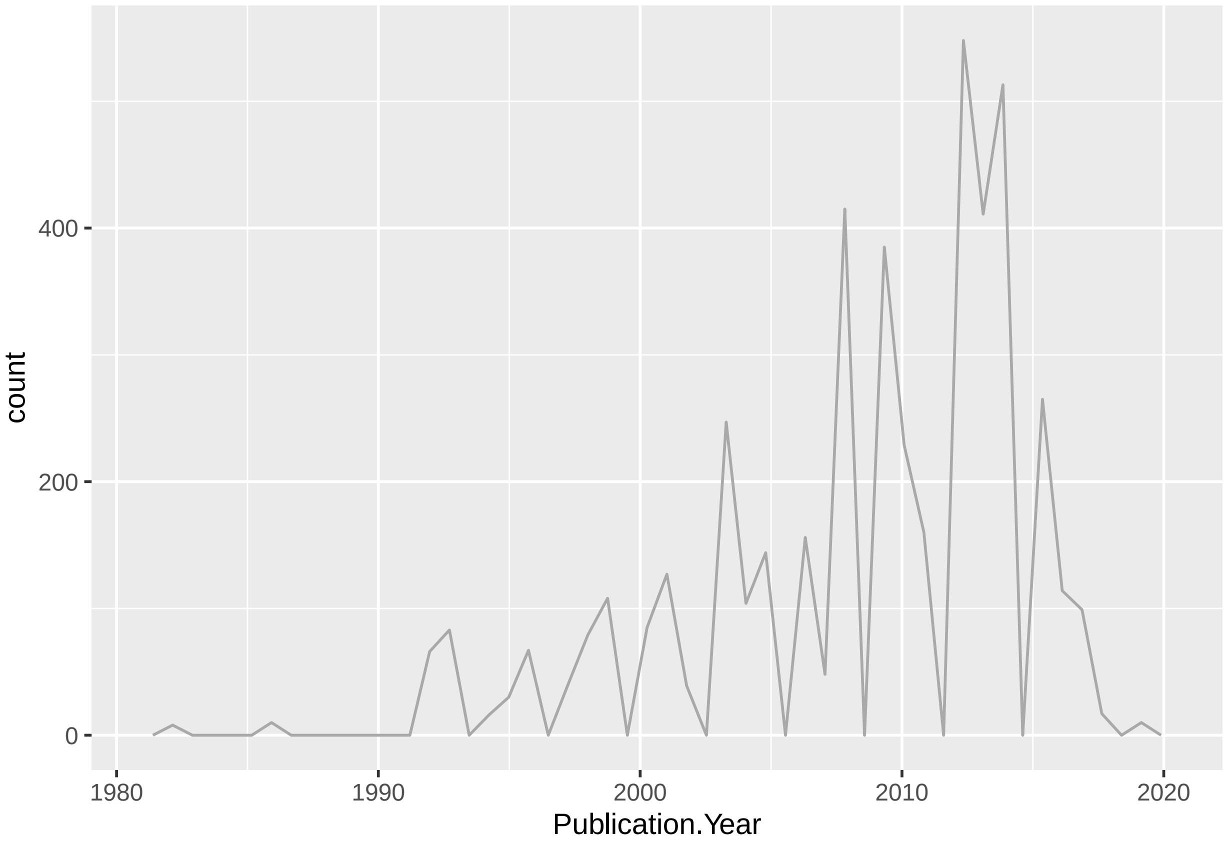
```
# class factor
```

Answer: This concentration variable is in the class factor. I’m guessing it’s factor and not numeric because the concentrations are discreet values in this case and are not continuous. There are also unusual notations in the variable, with some numbers written with a forward slash after them and a few cases where the variable is NR for not reported.

Explore your data graphically (Neonics)

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

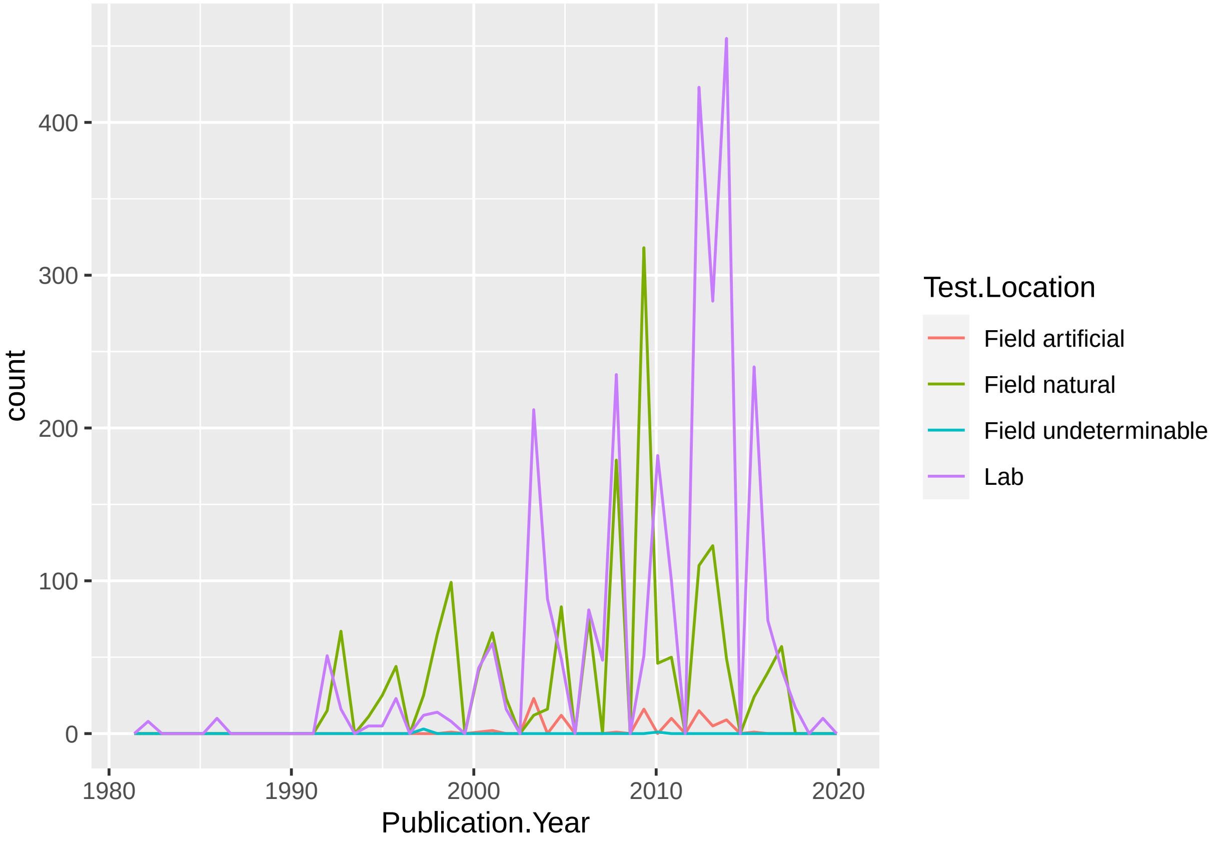
```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50, color = "darkgray")
```



```
# creates a plot showing the publication year on the x axis and a count of
# the number of studies published each year on the y axis
```

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 = 50)
```



```r
# creates a plot of publications counts per year with a line for each type of
# test location; a legend of test locations is generated to the right of the
# plot
```

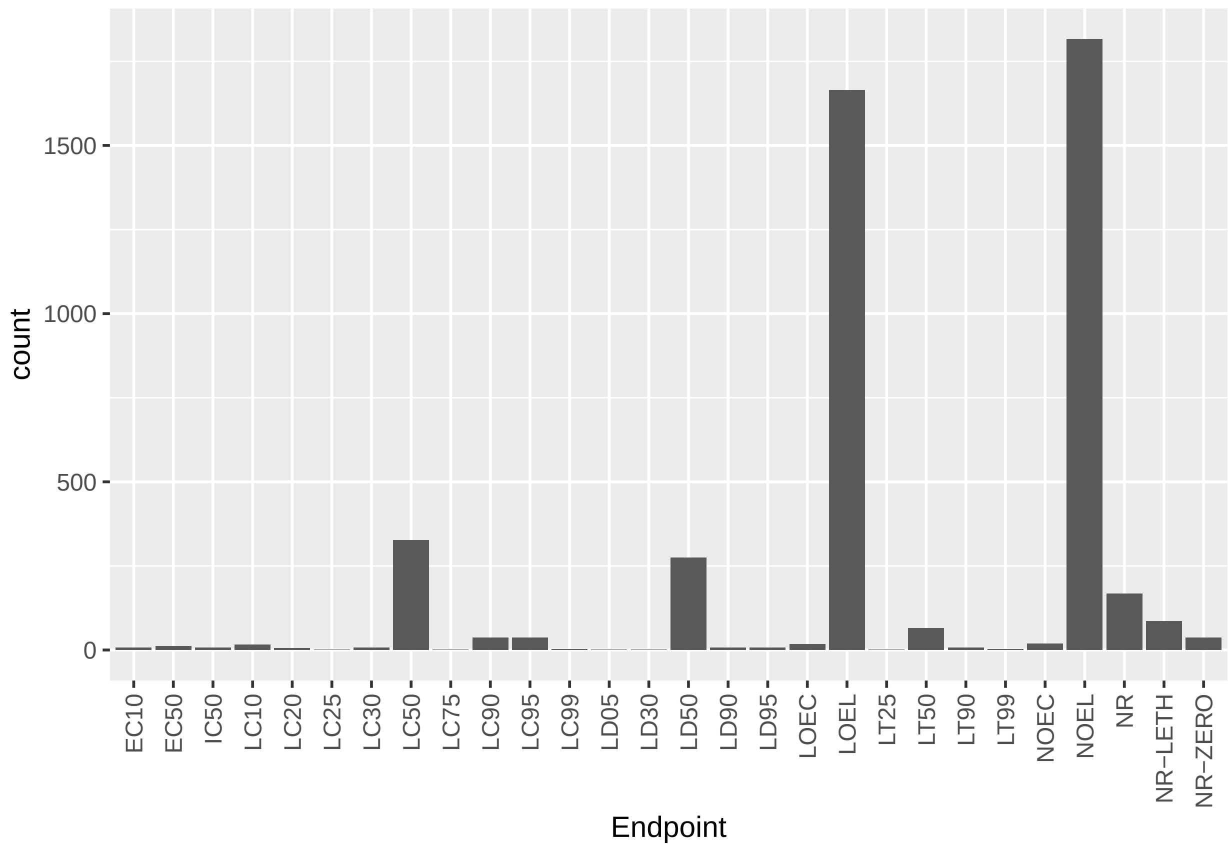
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are lab and field natural. It looks like they started out being used relatively equally, with maybe the field natural being slightly more popular. After 2010, the lab studies became by far the most popular test location. It looks like field artificial and field undeterminable were never that popular.

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
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



```r
# creates a barplot with endpoint on the x axis and the count of each on
# the y axis
```

Answer: The two most common endpoints are NOEL and LOEL. NOEL stands for no observed effect level and is the highest dose producing effects not significantly different from responses of controls (the highest does that is "safe"). LOEL stands for the lowest observed effect level, which means it was the lowest dose producing effects that were significantly different from responses of controls (the lowest dose that is harmful).

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
class(Litter$collectDate) # it is not a date it is a factor
```

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

```r
library(lubridate)
Litter$collectDate <- lubridate::ymd(Litter$collectDate) # convert to class date
# using lubridate

class(Litter$collectDate) # the class is now date
```

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

```
unique(Litter$collectDate) # sampling was done on 2018-08-02 and 2018-08-30
```

```
## [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$plotID) # there were 12 plots sampled at Niwot
```

```
##  [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) # so I can see the comparison of unique and summary
```

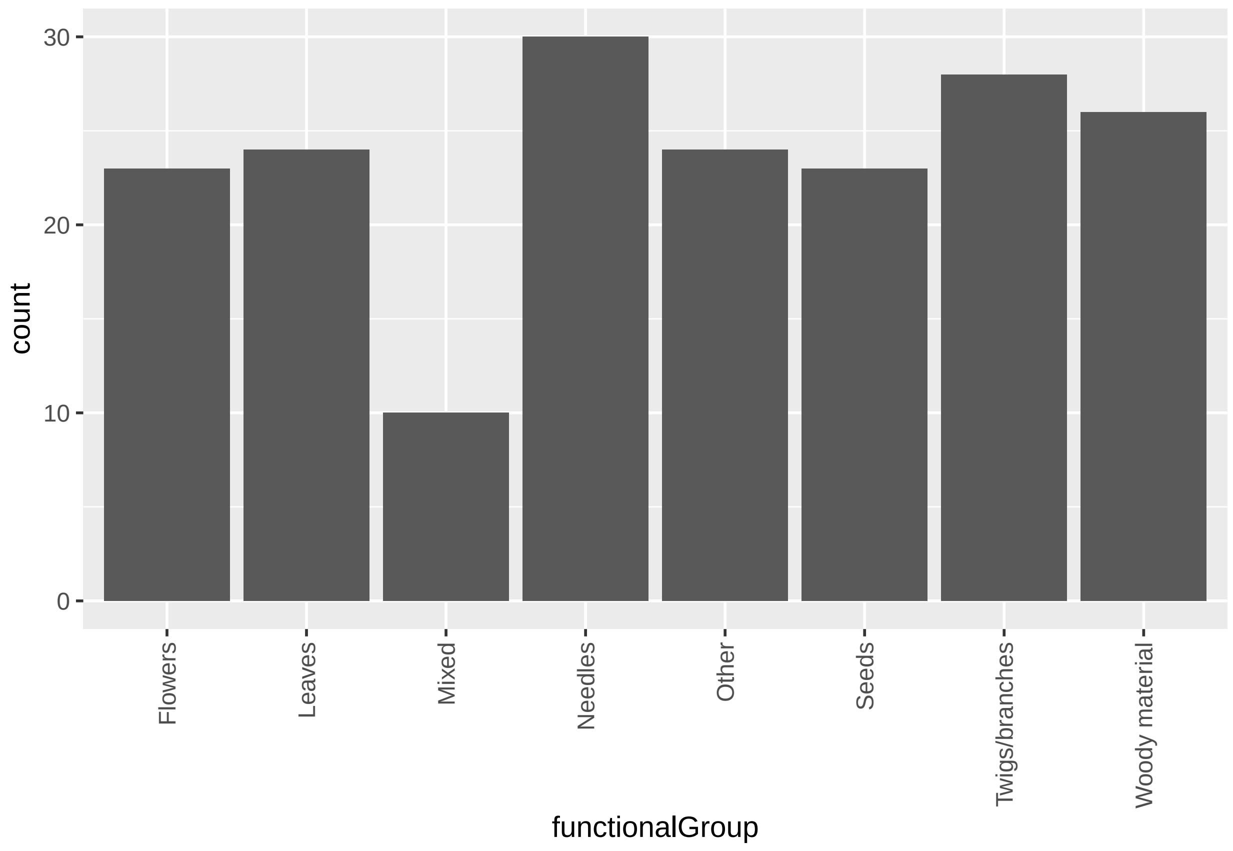
```
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061
##       20       19       18       15       14        8       16       17
## NIWO_062 NIWO_063 NIWO_064 NIWO_067
##       14       14       16       17
```

```
# outputs
```

Answer: The unique function provides a list of each of the plots that was sampled at the site. It's picking out each unique plotID regardless of how many of each plot are in the variable to show that there were 12 plots sampled at the site. The summary function then provides a count of how many of each of the plots are in the data set. So it provides the list of the 12 plots and then a count of how many times they're in the data set (for example, NIWO_040 has 20 entries).

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.

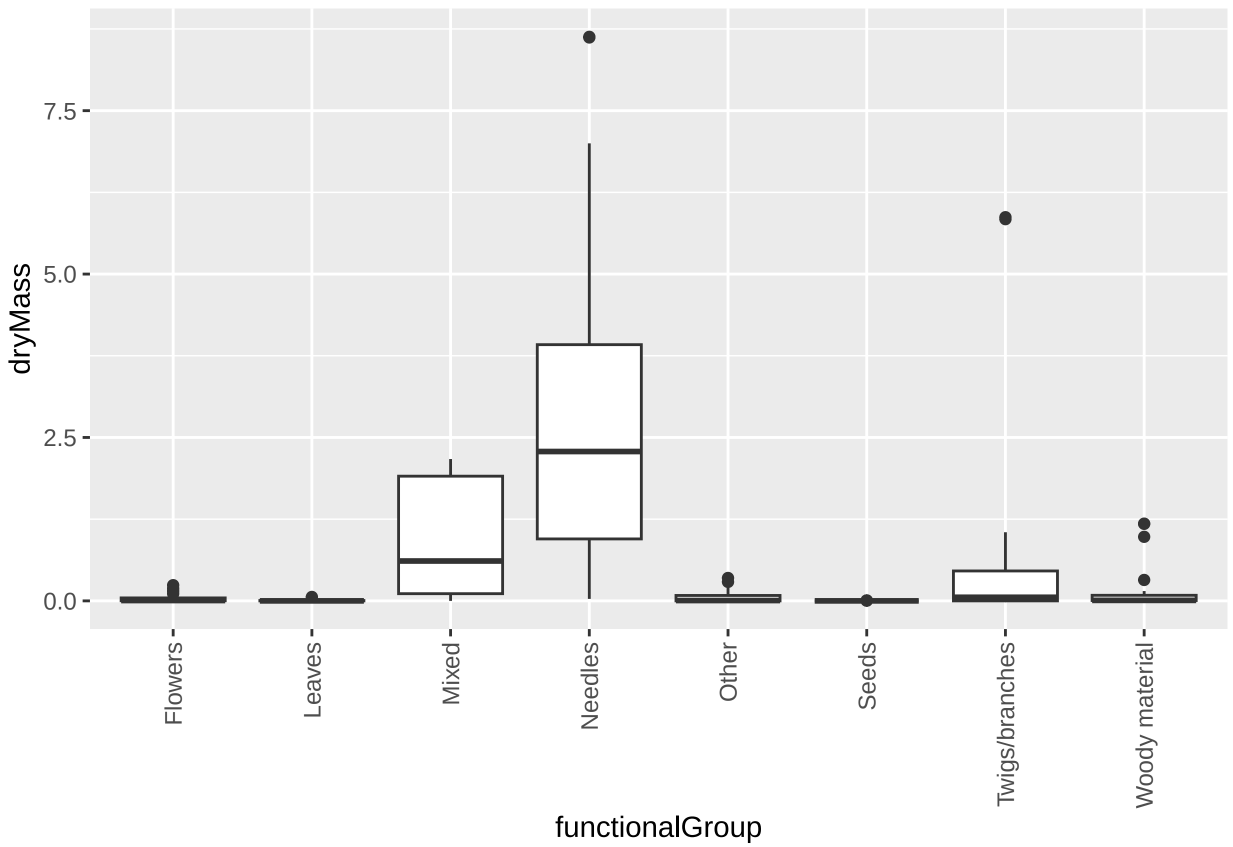
```
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



```
# creates a bar plot with the type of litter on the x axis and the count of
# each type on the y axis
```

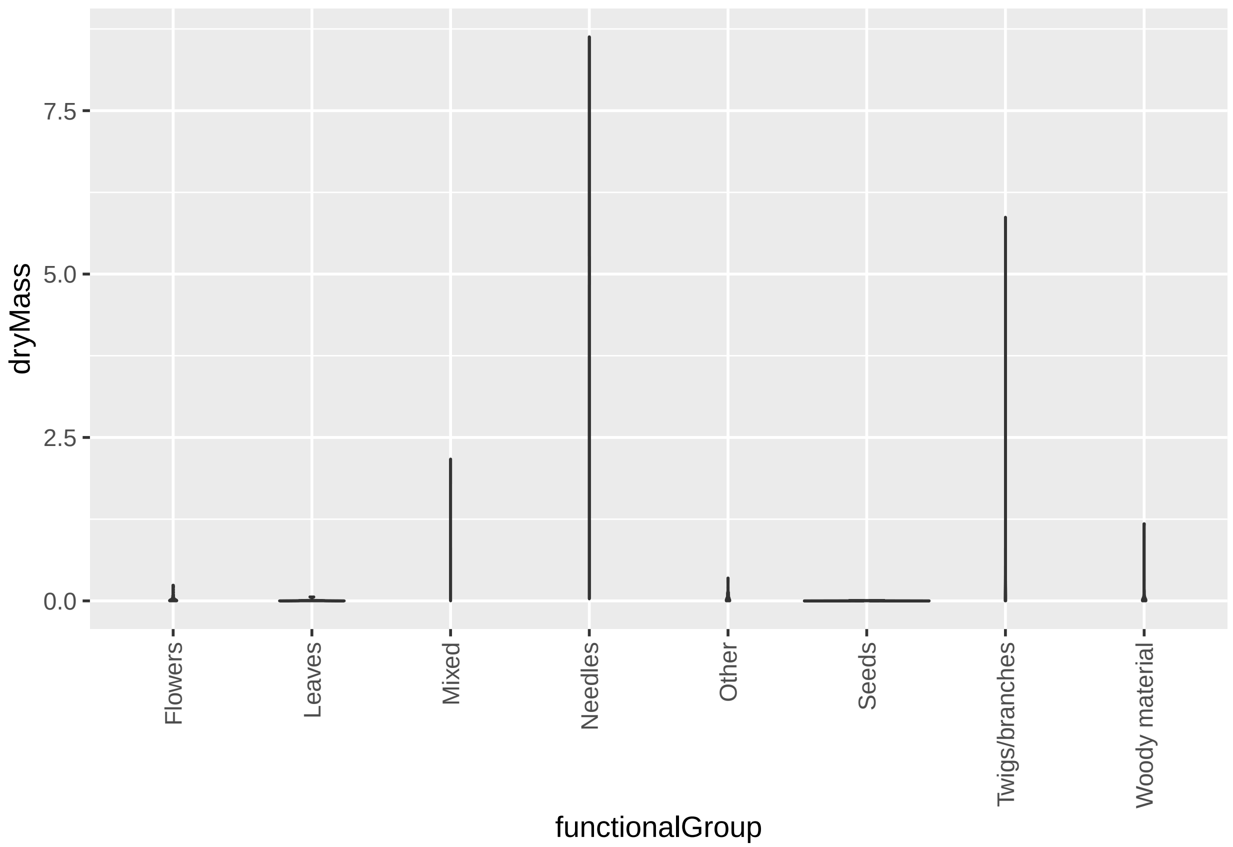
15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



```r
# generate the boxplot

ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



```
# generate the violin plot
```

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

> Answer: Violin plots display density distributions, but since there doesn't seem to be a large variation in density distribution in the types of litter with the largest dry masses, all you see are straight lines where you can't discern the summary statistics. Because the boxplot doesn't take into account the density distributions, you're able to see where the summary statistics are on the plot, making it more useful for this visualization.

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

> Answer: Needles, mixed, and twigs/branches are the types of litter which tend to have the highest biomass at these sites.