Assignment 3: Data Exploration

Annabelle White

Spring 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/ardwh/OneDrive/Documents/NSOE/env872/EDA-Spring2023/Assignments"
```

```
setwd("C:/Users/ardwh/OneDrive/Documents/NSOE/env872/EDA-Spring2023")
library(tidyverse)
library(lubridate) # Load packages
Neonics <- read.csv("Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = T)
Litter <- read.csv("Data/Raw/NIWO_Litter/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: Neonicotinoids are applied as seed treatments and may percolate through soil, groundwater, and nearby plants. If these chemicals prove toxic to various insect taxa, they could have widely spread deleterious impacts on decomposers, pollinators, and all other insects.

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: Litter and woody debris are crucial to cycling nutrients into forest soils via decomposition. They have a pronounced impact on carbon and nitrogen cycling.

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. Tower plot and litter trap distributions differ between forested (20 plots, random traps) and low-stature vegeration areas (30 plots, targeted traps). 2. Plot centers must be >50m from paved roads, and plot edges must be >10m from dirt roads. 3. Deciduous forests are sampled every 2 weeks, but discontinued for up to 6 months in the dormant season. Evergreen forests are sampled every 1-2 months.

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 are population (1803 studies) and mortality (1493 studies), several times more than the next most common effect (behavior). These make sense to study, as they can be surveyed through relatively uncomplicated methods, and give an overall snapshot of insect wellbeing. Trends in population and mortality would clearly indicate whether neonicotinoids are harming insect communities.

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...]

```
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
```

> Answer: The most commonly studied are species of bees and wasps. This makes sense, as bees and wasps are pollinators and provide an ecosystem service that is valuable to human food supply chains - including the agricultural fields that are employing neonicotinoids. If neonicotinoids are found to harm pollinators, this would have severe ramifications for their use.

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.)
```

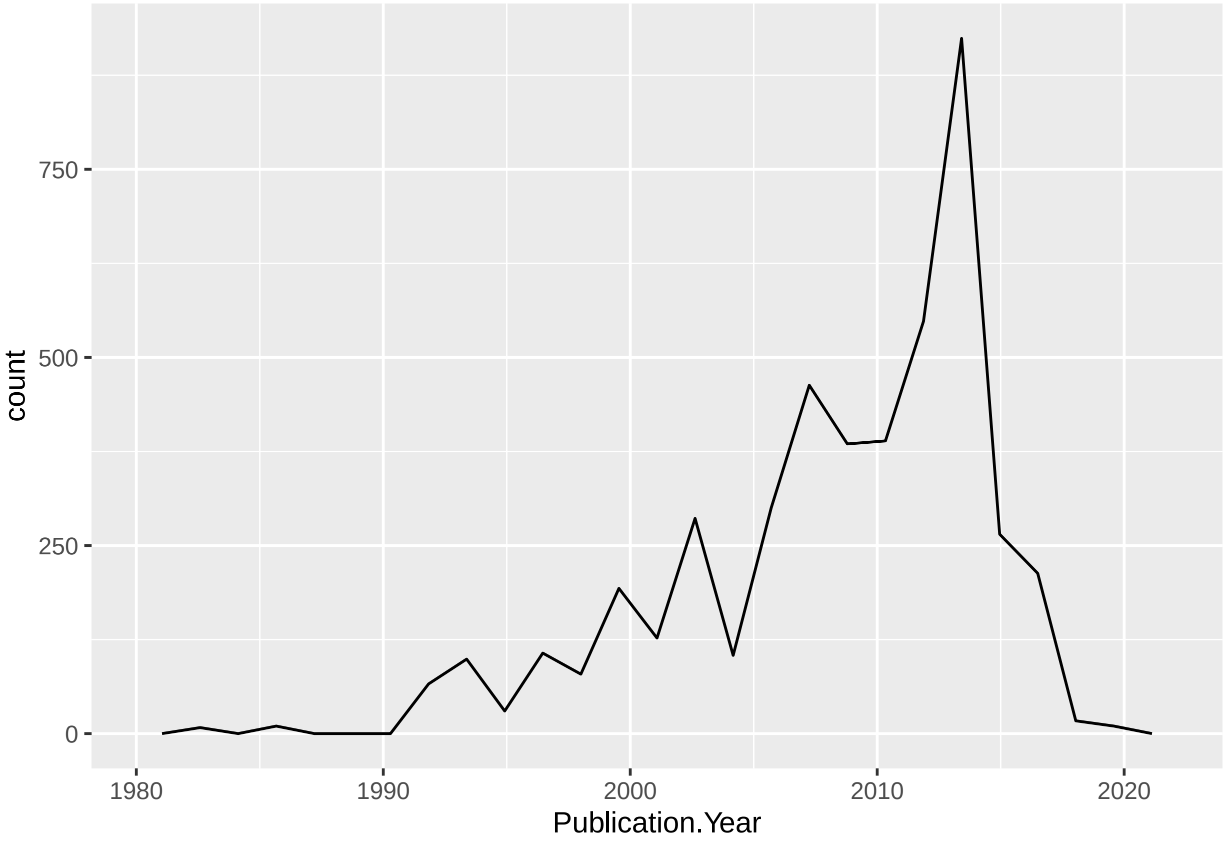
```
## [1] "factor"
```

> Answer: These data are not numeric because they contain characters such as >, /, and ~, which R does not process as numeric. This is good, because the data are not intended to be a continuous variable; they are arranged in categorical levels, comparing the differing levels of concentrations that each study used.

Explore your data graphically (Neonics)

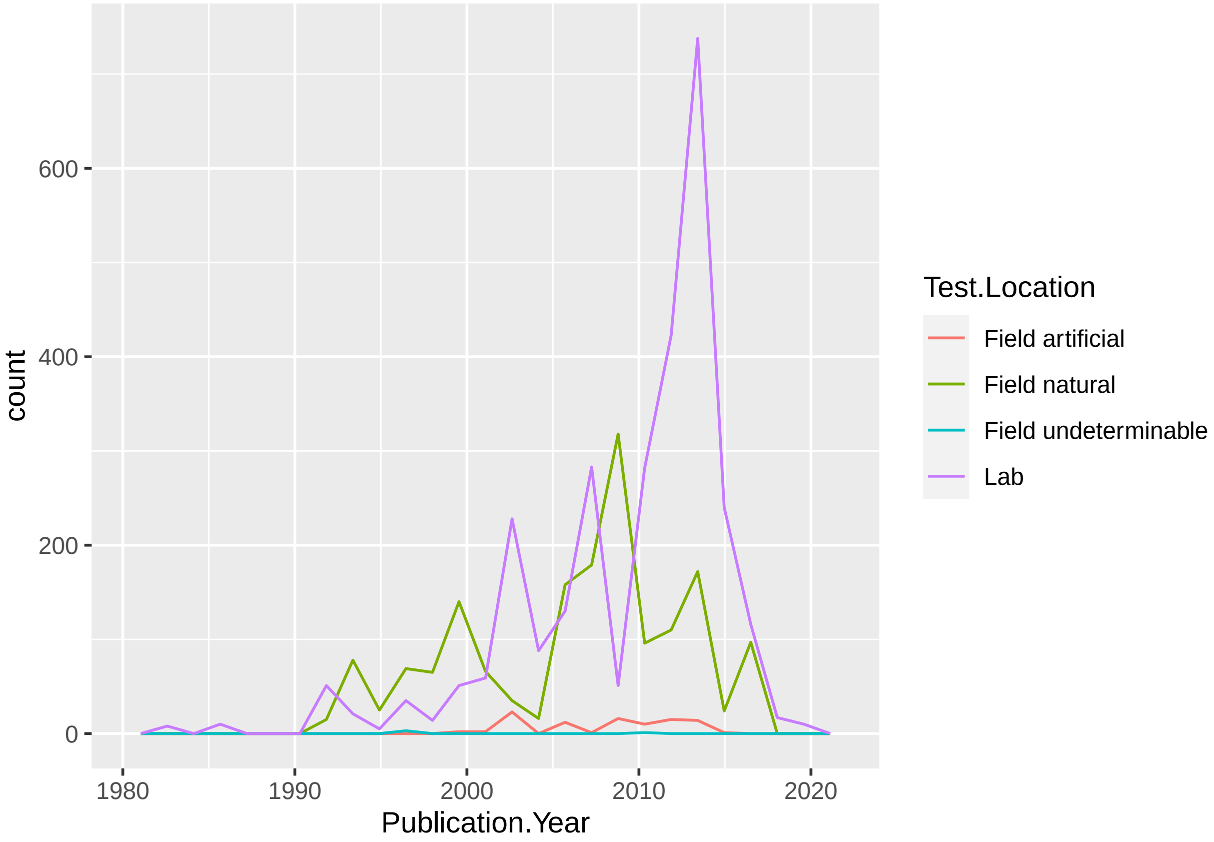
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 25) # Looks nicer this way
```



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 = 25)
```



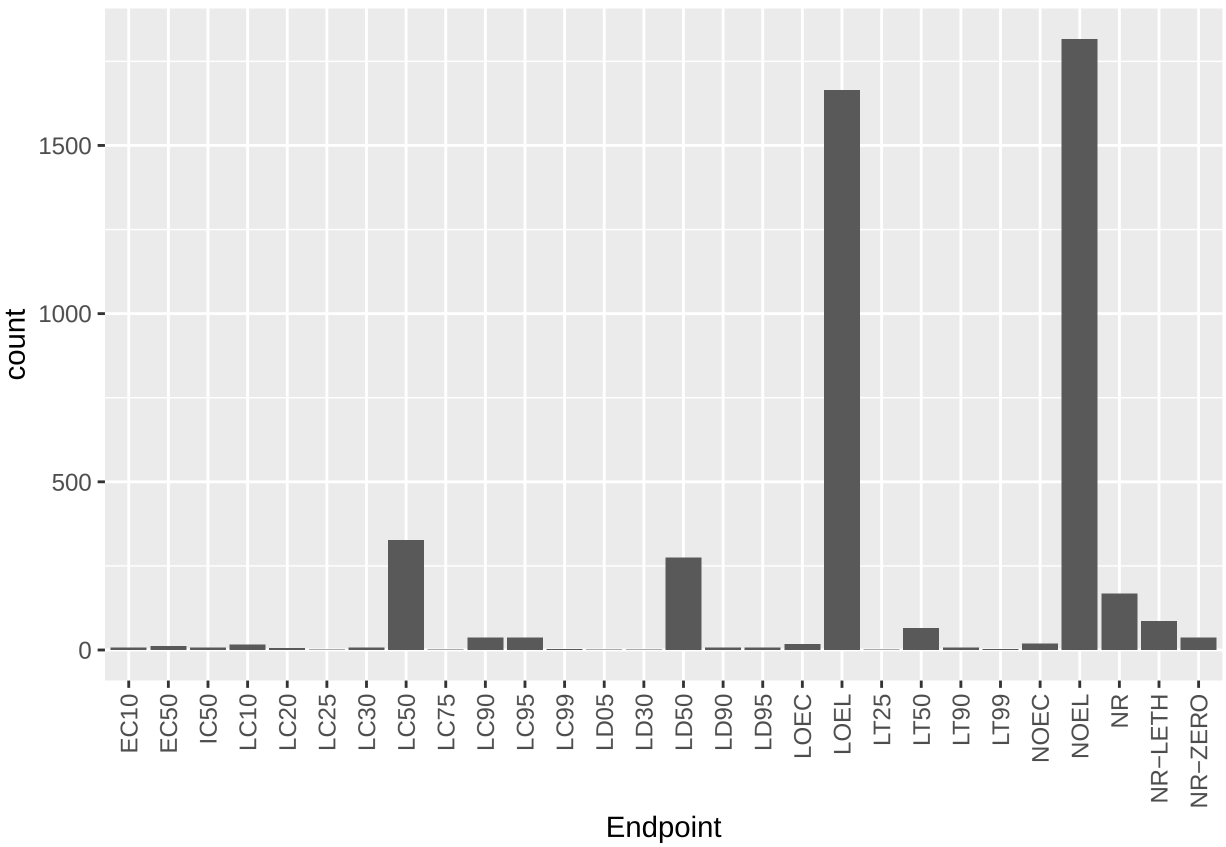
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are in natural field settings and lab settings. Which of these is most common varies, with spikes in certain periods. For example, in the 2010s, lab tests increased to drastically more than any other test location, comprising most of the tests.

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



Answer: The most common endpoints are Lowest Observed Effect Level (LOEL), in which the lowest dose of concentration produced effects significantly different from control; and No Observed Effect Level (NOEL), in which the highest dose of concentration did not produce such effects.

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.

```
# Fix data type
class(Litter$collectDate) # Not a date; displays as factor (originally)
```

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

```
summary(Litter$collectDate) # To show the date format; it's year-month-day
```

```
## 2018-08-02 2018-08-30
##         91         97
```

```
Litter$collectDate <- ymd(Litter$collectDate) # Perform Y/M/D function on the vector
class(Litter$collectDate) # Check class; it's a date!
```

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

```
# Check for all unique dates:
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`?

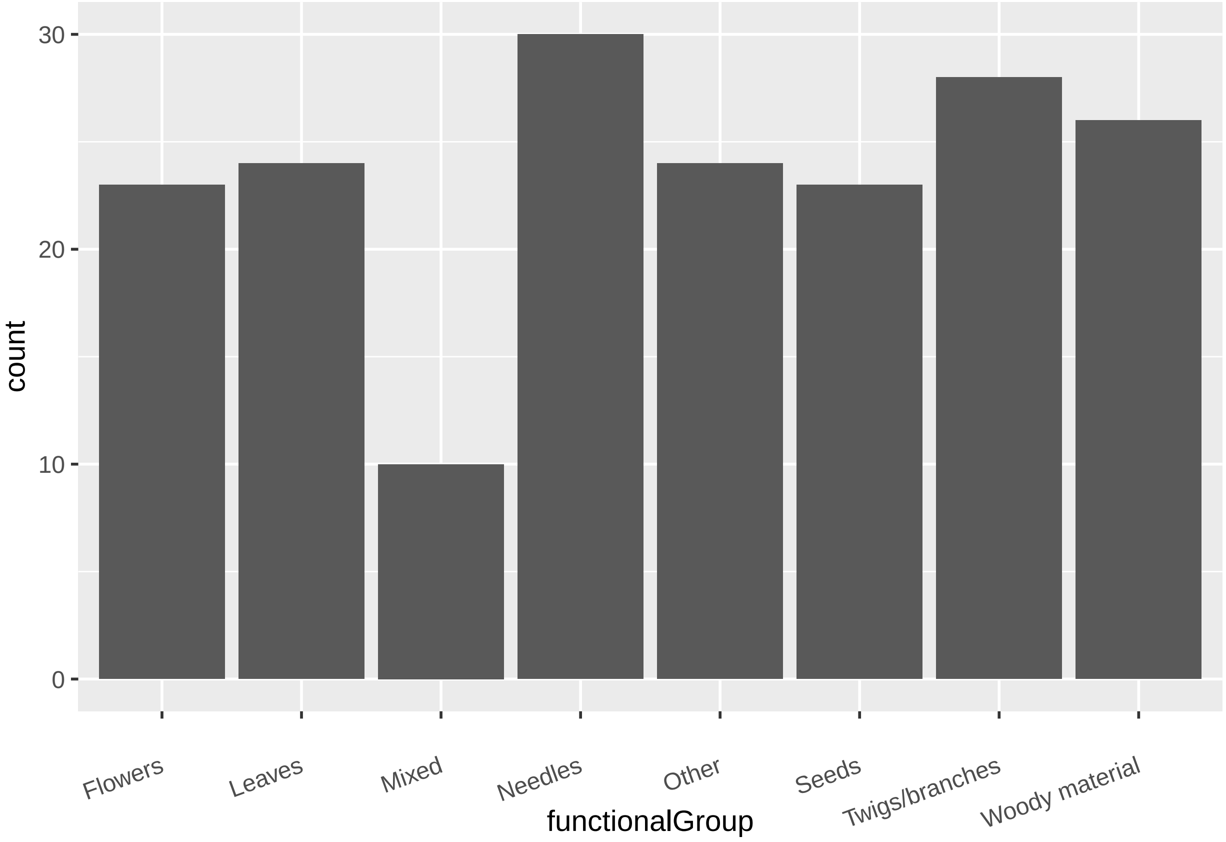
```
unique(Litter$plotID)
```

```
##  [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
```

Answer: summary() shows the distribution of values in a vector. unique() shows the number of distinct values in a vector. You could get the same information from summary() if you manually counted the display, but it seems foolish to do that in a coding course, and utterly unfeasible for larger datasets.

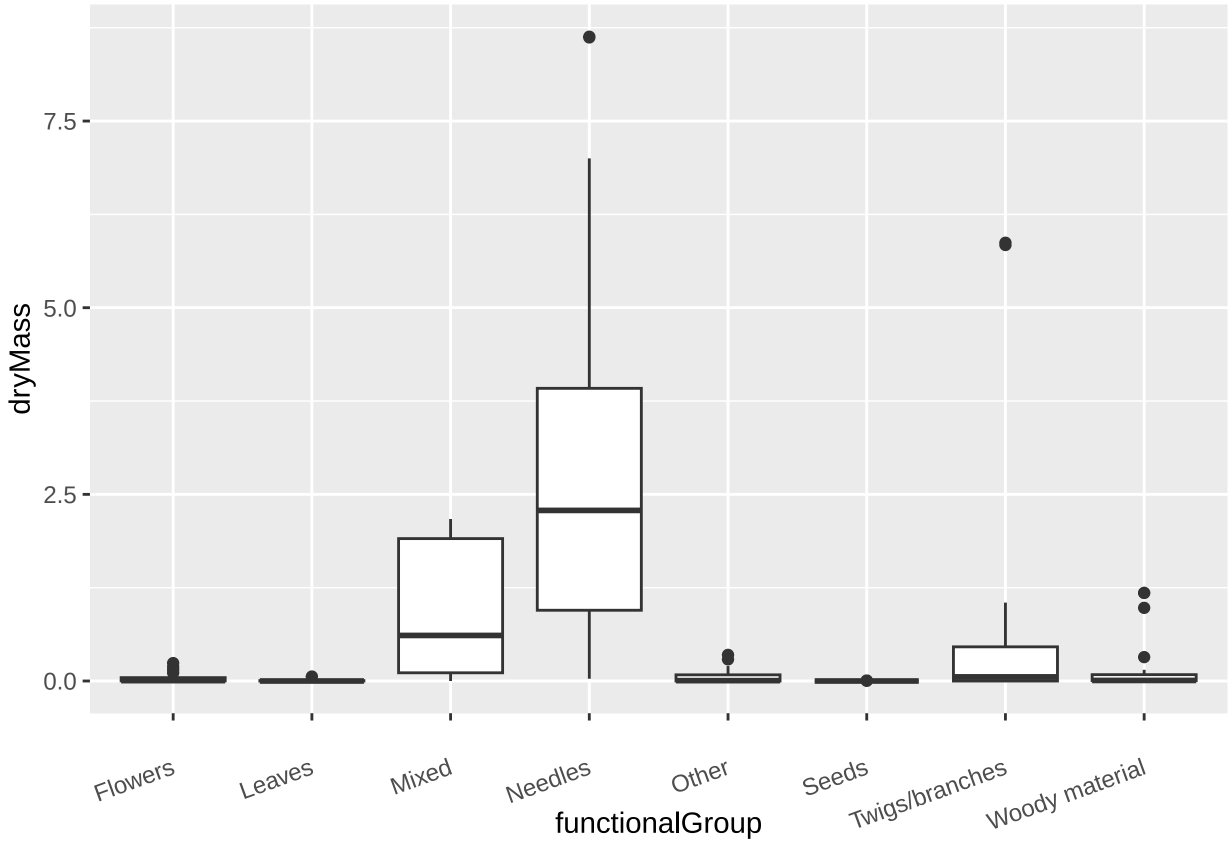
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter) +
    geom_bar(aes(x = functionalGroup)) +
  theme(axis.text.x = element_text(angle = 20, vjust = 0.5, hjust=1)) # Tilt labels
```

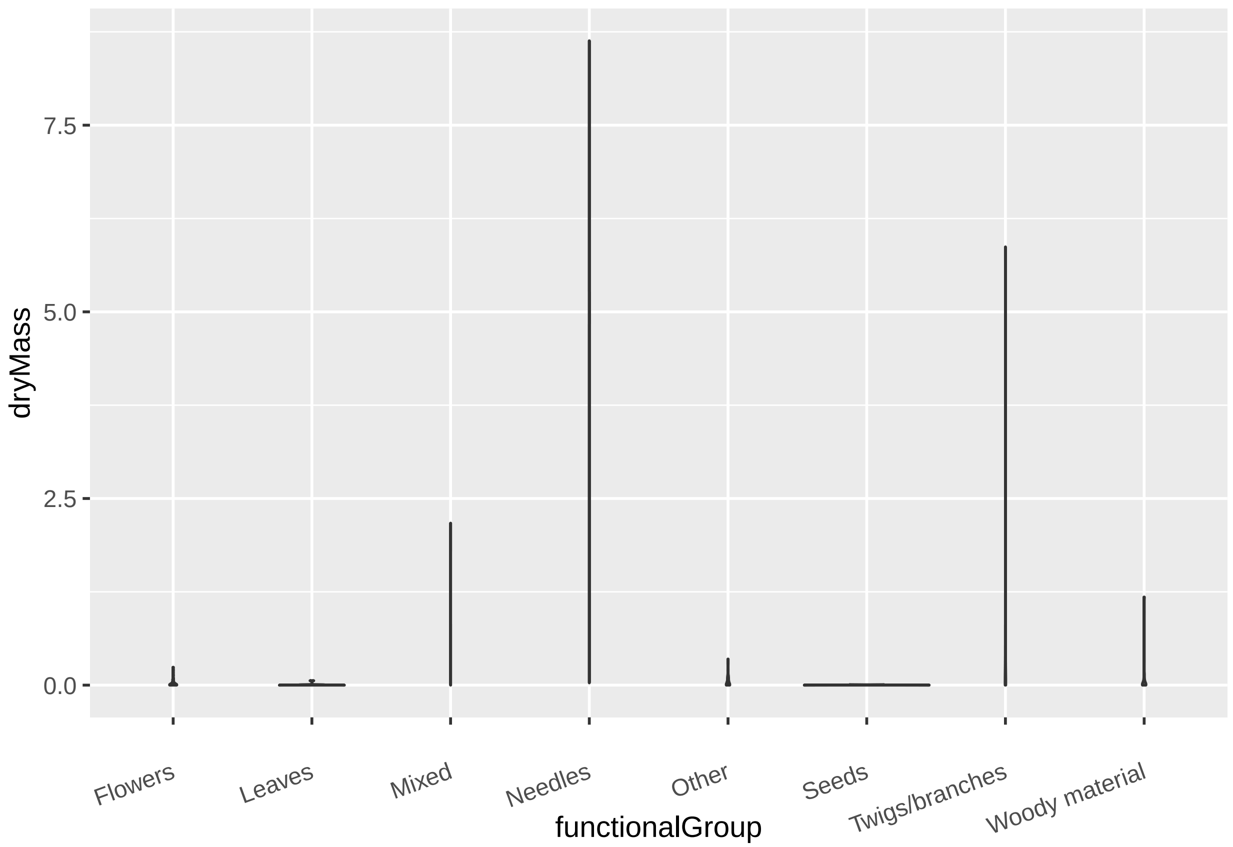


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(Litter) +
    geom_boxplot(aes(x = functionalGroup, y = dryMass, group = cut_width(functionalGroup, 1))) +
  theme(axis.text.x = element_text(angle = 20, vjust = 0.5, hjust=1)) # we slidin
```



```
ggplot(Litter) +
    geom_violin(aes(x = functionalGroup, y = dryMass)) +
  theme(axis.text.x = element_text(angle = 20, vjust = 0.5, hjust=1))
```



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

> Answer: Because there is a very even distribution of masses across most functional groups, there is little variation to be displayed on the violin plot. As a result, the violins have been gruesomely flattened into single strings. As any orchesta concertgoer can attest, it is much more difficult to interpret the outputs from a single string than from a healthy violin. The boxplot is more effective because it shows that there is very little variation in mass for most functional groups.

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

> Answer: The highest biomass comes from needle litter and, to a lesser extent, mixed litter. This makes intuitive sense, as anyone who's ever set foot in a pine stand can tell you that litter is DENSE. There is a spot off a trail in the Duke Forest that I think would be a perfect place to take a nap in the dappled mid-morning sun. I crave it.