Assignment 3: Data Exploration

Iman Byndloss

Fall 2024

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 Canvas.

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. Load necessary packages (tidyverse, lubridate, here), check your current working directory 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.

```r
# Load all necessary packages, using library().
# If not already installed, install() in the console.
library(tidyverse); library(lubridate); library(here)

# Check working directory, using here().
here()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```r
# Upload all necessary data, using read.csv().
# Name them as described and ensure that strings are read as factors.
Neonics <- read.csv(
  file = here('Data','Raw','ECOTOX_Neonicotinoids_Insects_raw.csv'),
  stringsAsFactors = TRUE)
Litter <- read.csv(
  file = here('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: It is not surprising that there is substantial research on how neonicotinoids affect insects, especially given there are likely important questions regarding its efficacy, duration, and unintended consequences. The answers to such questions could influence how people use neonicotinoids.

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: Researching forest litter and woody debris is likely valuable to determining the influence such material has on the surrounding environment. Potentially important questions could surround its duration, changing chemical composition, and ecosystem roles.

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. Spatial Sampling Design: Sampling occured at terrestrial NEON sites that contain woody vegetation greater than 2 meters tall and in randomly selected tower plots (within 90% flux footprint of the primary and secondary airsheds). 2. Spatial Sampling Design: Trap placement within plots depended on vegetation. For example, in sites with more than 50% aerial cover of woody vegetation greater than 2 meters in height, traps were randomly placed, using a randomized list of grid cell locations. 3. Temporal Sampling Design: Although ground traps are sampled once per year, elevated traps vary based on the surrounding vegetation, with biweekly sampling in deciduous forest sites and bimonthly sampling in evergreen forest sites.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
# Used dim() to retrieve the dimensions of the Neonics dataframe.
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? [Tip: The `sort()` command is useful for listing the values in order of magnitude...]

```
# Used summary() to produce a results summary for "Effect" in the Neonics dataframe.
# This was done within sort() to list of values was in order of magnitude (least to greatest).
sort(summary(Neonics$Effect))
```

```
##       Hormone(s)        Histology       Physiology          Cell(s) 
##                1                5                7                9 
##     Biochemistry     Accumulation     Intoxication    Immunological 
##               11               12               12               16 
##       Morphology           Growth        Enzyme(s)         Genetics 
##               22               38               62               82 
##        Avoidance      Development     Reproduction Feeding behavior 
##              102              136              197              255 
##         Behavior        Mortality       Population 
##              360             1493             1803
```

Answer: Population (1803) and Mortality (1493) are by far the most common effects of interest, which is likely due to the commonality of research attempting to determine how population numbers of insects are directly impacted by neonicotinoids. Such information can help determine how effective the insecticide is overall and for distinct species.

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: Explore the help on the `summary()` function, in particular the `maxsum` argument...]

```
# Used summary() to produce a results summary for "Species.Common.Name" in the Neonics dataframe.
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
```

```
# help(summary) describes maxsum;
# "integer, indicating how many levels should be shown for factors".
# With the last group being "(Other)", inputted 7 as the integer to ensure 6 species were displayed.
summary(Neonics$Species.Common.Name, maxsum=7)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee      Italian Honeybee
##                   152                   140                   113
##               (Other)
##                  3083
```

Answer: The six most commonly studies species include the Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee (in that order). All of these species are members of the Hymenoptera order and act as pollinators. These species are likely non-target species for the insecticide, which is probably why they are more often studied, along with the fact that the decline of such species can have cascading effects for ecosystems.

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? [Tip: Viewing the dataframe may be helpful…]

```
# Used class() to determine the class for "Conc.1..Author." in the Neonics dataframe.
# Used view() to view the dataframe.
class(Neonics$Conc.1..Author.)
```

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

```
# view(Neonics)
```

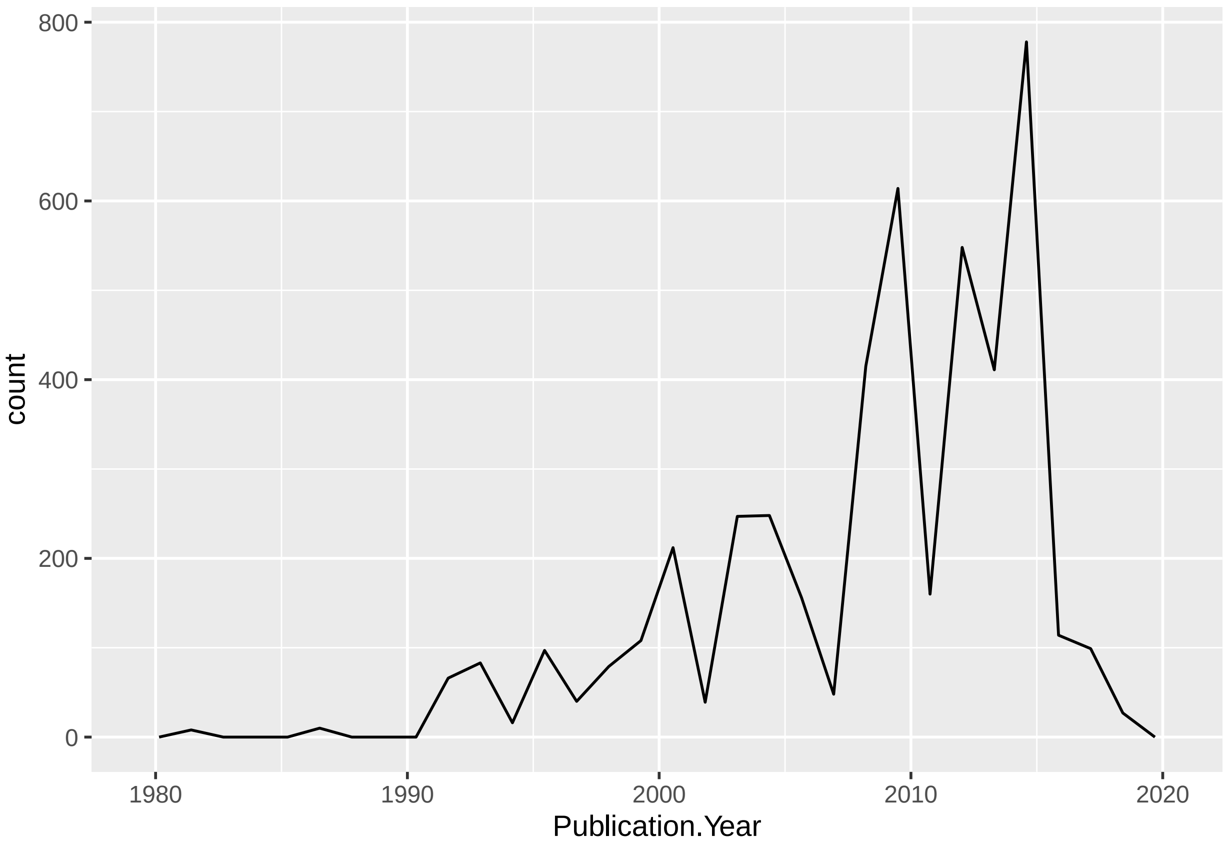
> Answer: The “Con.1..Author” is classified as factor, not numeric. This classification could be due to certain values containing less than symbols, leading values to be interpreted as being in categories.

Explore your data graphically (Neonics)

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

```
# Typed help("geom_freqpoly") in the console to learn how to use the function.
# Looked back at 03_DataExploration_Part2.Rmd.
# Used geom_freqpoly() to generate a frequency plot;
# number of studies versus publication year for the Neonic dataframe.
ggplot(Neonics) +
  geom_freqpoly(aes(x=Publication.Year))
```

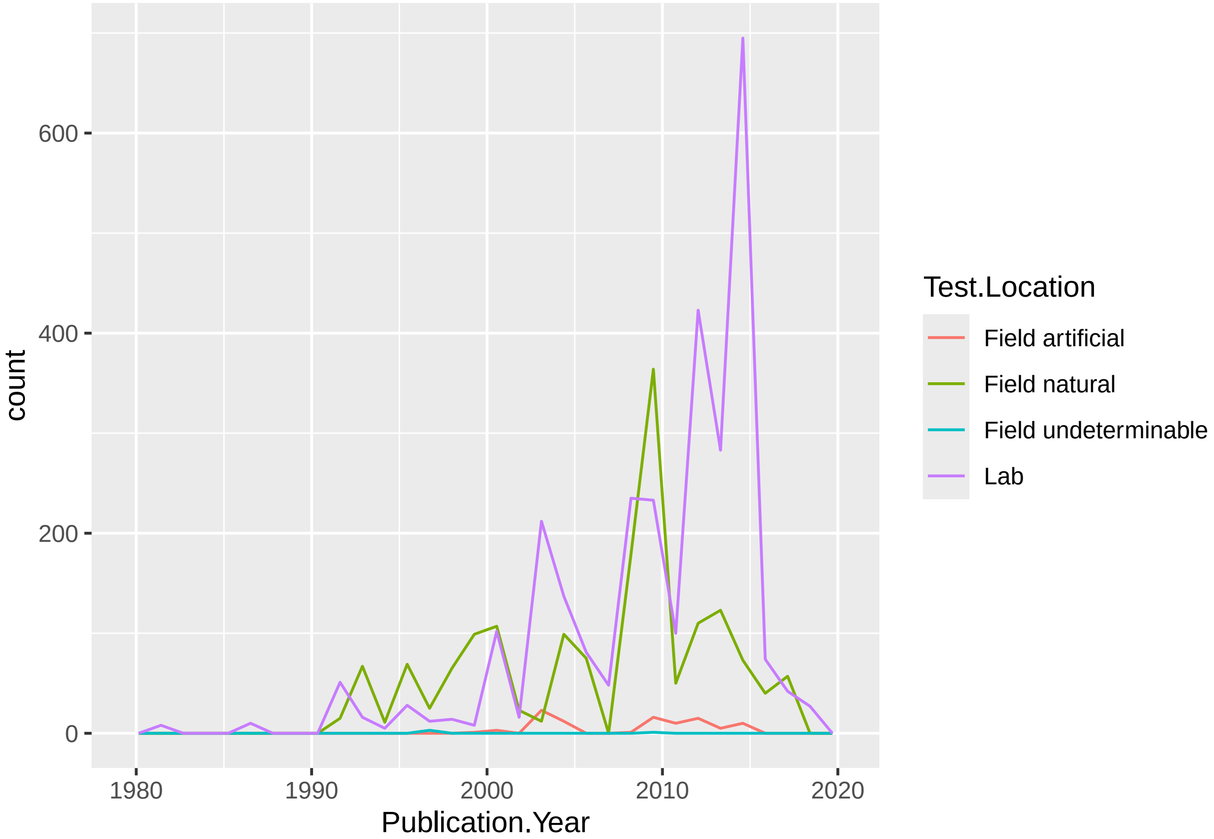
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



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

```
# Based on the previous code, added a line within aes();
# it specifies color will be determined by Test.Location.
ggplot(Neonics) +
  geom_freqpoly(aes(x=Publication.Year, color=Test.Location))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



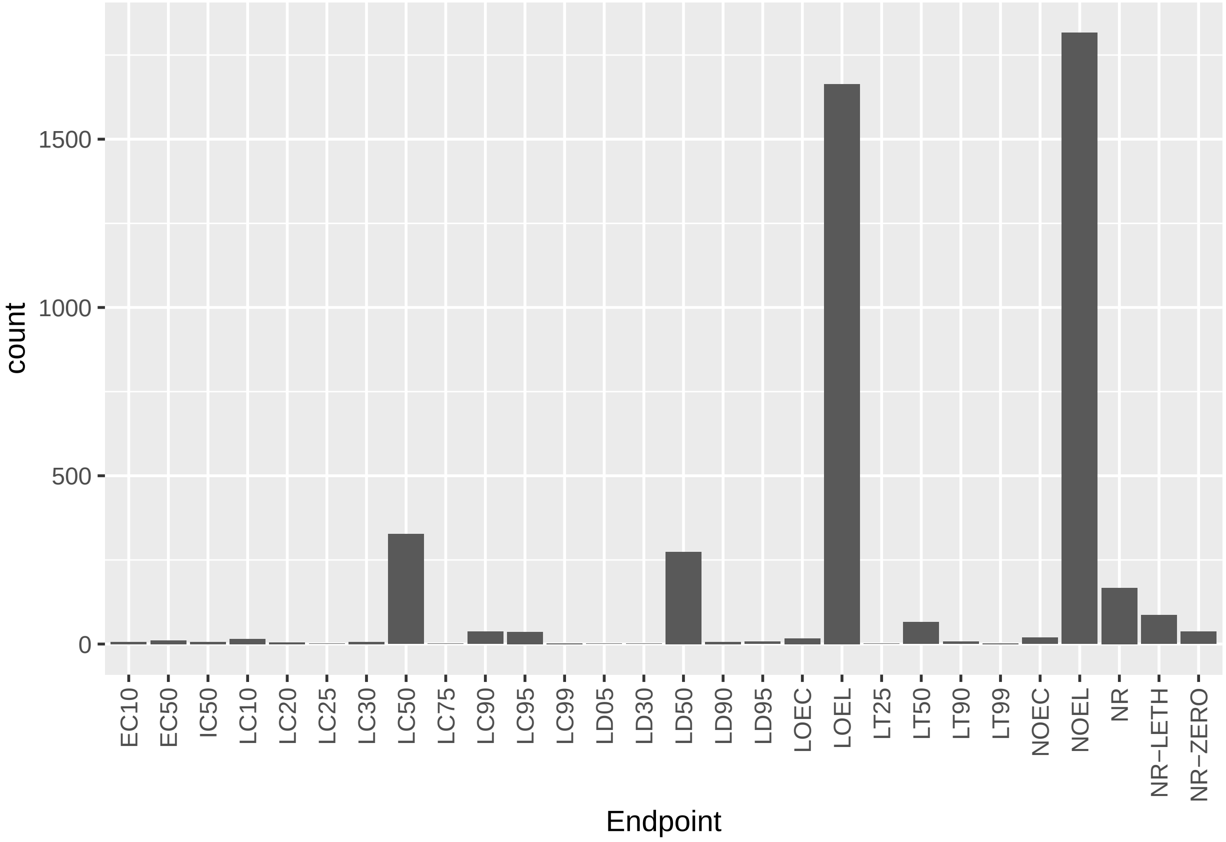
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The two most common test locations by far are lab and field natural. Lab and field natural switch with one another for the top spot. For example in 2010, field natural was the most common, but in 2015, lab was the most common after drastically increasing in frequency in more recent years.

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…]

```
# Used geom_bar() within ggplot() to create a bar graph;
# "Endpoint" in the Neonics dataframe.
# Added the suggested line above.
ggplot(Neonics) +
  geom_bar(aes(x=Endpoint)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: The two most common endpoints by far are NOEL and LOEL, respectively. While both these enpoints are associated with terrestrial database usage, NOEL is defined as no-observable-effect-level (i.e., highest dose producing effects not significantly different from responses of controls) and LOEL is defined as lowest-observable-effect-level (i.e., lowest dose producing effects that were significantly different from responses of 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.

```
# Used class() to determine the class for the "collectDate" in the Litter dataframe.
# Used view() to view the dataframe.
class(Litter$collectDate)
```

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

```
# Looked at 03_DataExploration_Part1.Rmd for reference.
# Since it returned as a factor, used as.Date() to establish it as a date with the format %Y-%m-%d.
# The four digit year corresponds to "%Y", not "%y".
Litter$collectDate <- as.Date(Litter$collectDate, format="%Y-%m-%d")

# Used unique() to determine which dates litter was sampled during August 2018.
# Within the code, specified 1) the format, 2) the condition, and 3) what to return.
aug2018litter <- unique(Litter[format(Litter$collectDate, "%Y-%m") == "2018-08", "collectDate"] )
```

13. Using the `unique` function, determine how many different plots were sampled at Niwot Ridge. How is the information obtained from `unique` different from that obtained from `summary`?

```
# Used summary() for comparison to the output of the following code.
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
```

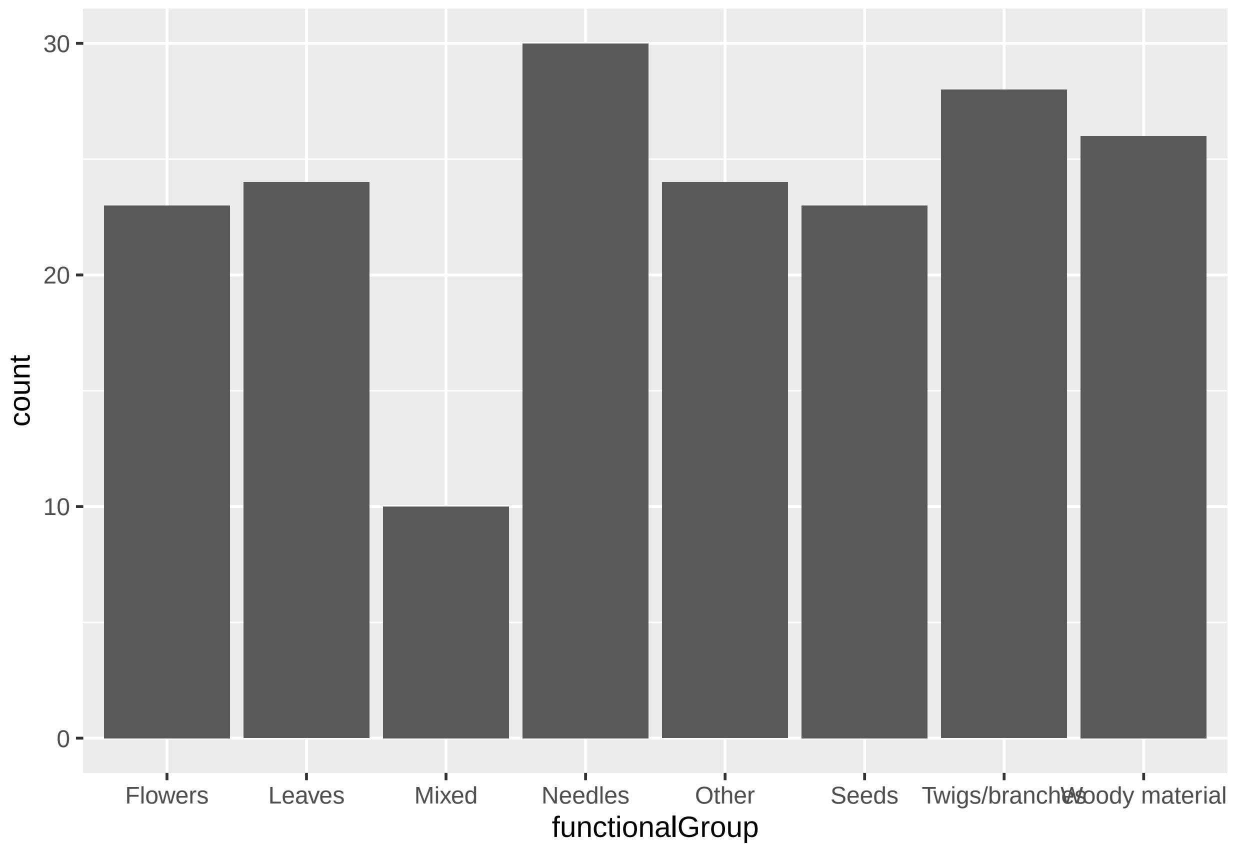
```
# Used unique() to determine the different plots.
# Put it within length() to ensure that those plots were counted.
length(unique(Litter$plotID))
```

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

Answer: While length(unique()) gives the count for the number of unique variables in the "plotID" column (i.e., 12), summary() gives the total counts for each of the 12 unique variables in the "plotID" column.

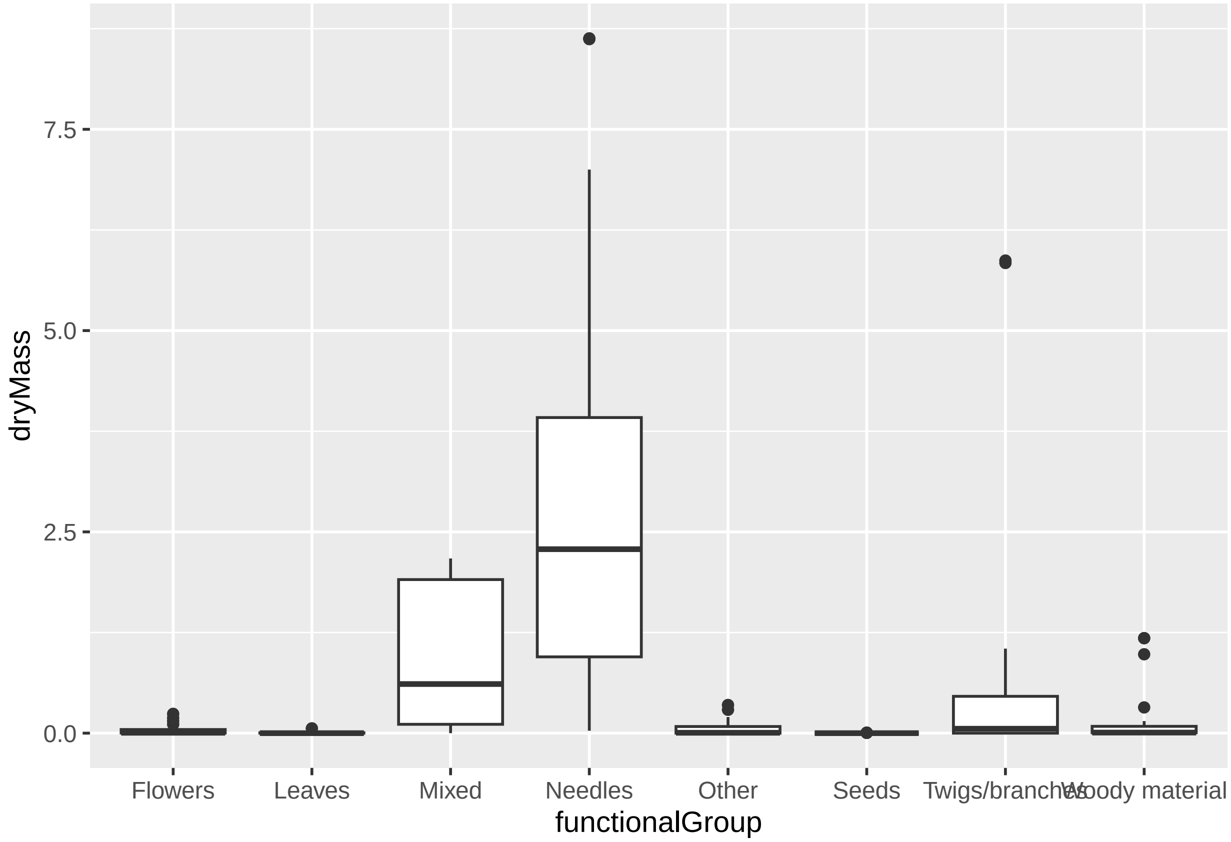
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
# Used geom_bar() within ggplot() to create a bar graph;
# "functionalGroup" in the Litter dataframe.
ggplot(Litter) +
  geom_bar(aes(x=functionalGroup))
```

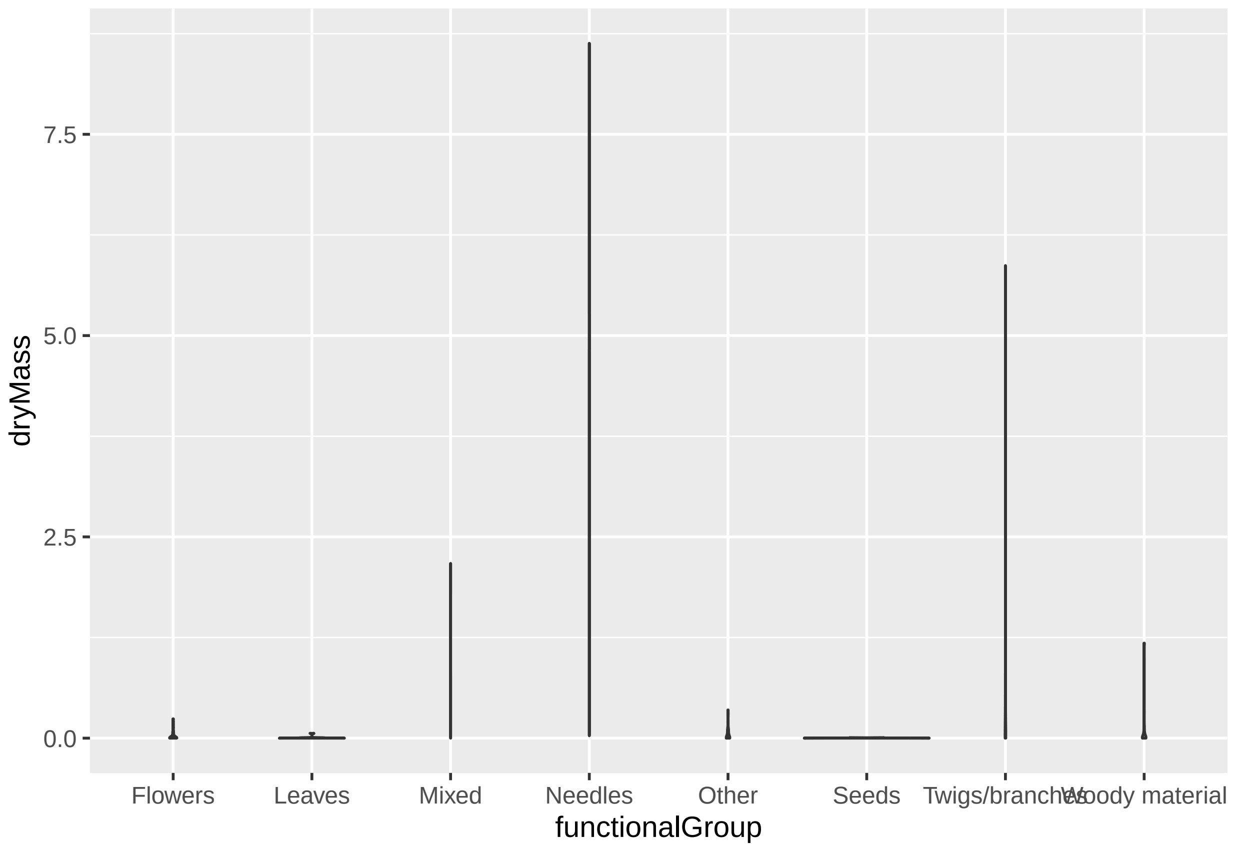


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
# Used geom_boxplot() within ggplot() to create a boxplot;
# "functionalGroup" vs "dryMass" for the Litter dataframe.
ggplot(Litter) +
  geom_boxplot(aes(x=functionalGroup, y=dryMass))
```



```r
# Used geom_violin() within ggplot() to create a violin plot;
# "functionalGroup" vs "dryMass" for the Litter dataframe.
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: Both graphing methods depict distributions of numeric data for one or more groups, but a boxplot uses boxes and lines while a violin plot uses density curves, combining a boxplot with a density plot. In this case, the boxplot is a more effective visualization option than the violin plot, as the violin plot is unable to display any useful density information.

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

> Answer: The four types of litter with the highest biomass at these sites are Needles, Twigs/branches, Mixed, and Woody material (in that order), with all other functional groups being relatively minimal. Still, Needles and Twigs/branches are more individually than all other groups combined.