Assignment 3: Data Exploration

Kallie Davis

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.

```
#install.packages("formatR")
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=80), tidy=TRUE)
```

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.

```
getwd()
```

```
## [1] "C:/Users/kalli/OneDrive/Desktop/Grad_School/Environmental Data Analytics_ENV_872/EDA-Fall2022/As
```

```
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: When applying an insecticide it is important to understand how it affects not only the target insect species but also other species in the insect community. If you use an insecticide you want the dose and specific metabolic targets to have the largest impact on the specific insect of interest while ideally having little to no effect on non-target species. This, however, is a very challenging task. Insecticides traditionally target a specific metabolic process for insects with (ideally) little to no effect on higher order organisms. This approach still leaves other local insects, such as bees, susceptible to its application.

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 a reflection of nutirent cycling in natural systems; they provide nutirients back to the soil, stream, or other systems which can be broken down by microbes and uptaken again by living organisms. Studying litter production can serve as an indicator of primary productivity in a forest ecosystem. The accumulation of litter or woody debris in forests can also lead to increased risk for local fires with a changing climate which is a large concern in the Colorado region.

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. litter is defined as material which is less than 2 centimeters (cm) in butt-end diameter, is less than 50 cm long, and fell from the forest canopy. Wood debris is defined as material which is greater than 2 centimeters (cm) in butt-end diameter, is greater than 50 cm long, and fell from the forest canopy. 2. Samples are taken both in ground and elevated traps. The number of sampling plots is dependent on the tower airshed size and traps are placed in either a targeted or randomized fashion. 3. Sampling at ground traps occurs once every year; sampling of elevated traps varys depending on the site vegetation.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)  #The Neonics data has 4623 rows and 30 columns
```

```
## [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: Effects on mortality, population, and behavior are the most commonly studied effects. With the use of an insecticide one is often interested in mortality as a function of effectiveness. Mortality naturally has effects on the population which also makes this an important factor to study. Insecticides can also have other undesireable effects on behavior which can also cause adverse population survival effects of target and nontarget 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.

```
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: Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, Italian Honeybee are the most commonly studies species. They are all important pollinator species, thus making it important to understand the adverse effects of insecticides on them as this will also laregly effect the local ecosystems.

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"
```

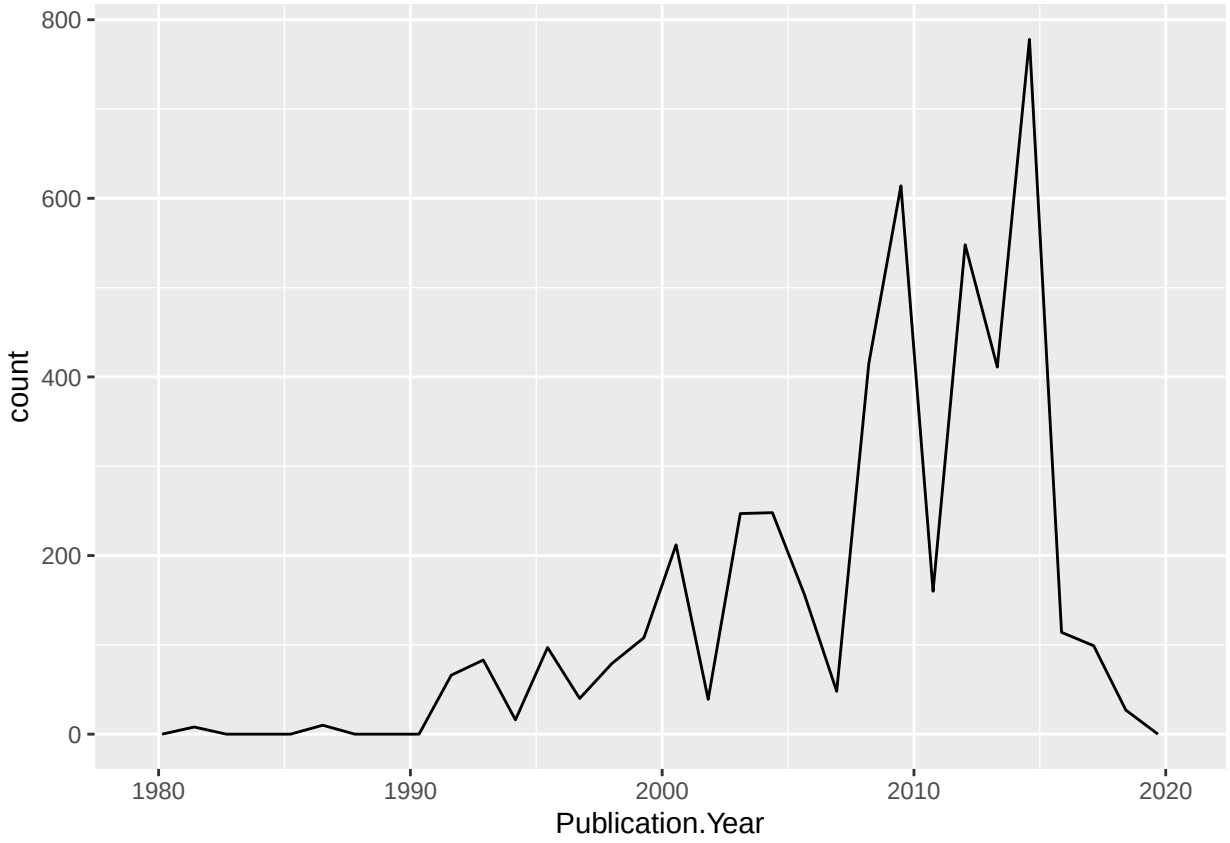
Answer: when reading the csv file I made it so strings would be seen as factors.

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))
```

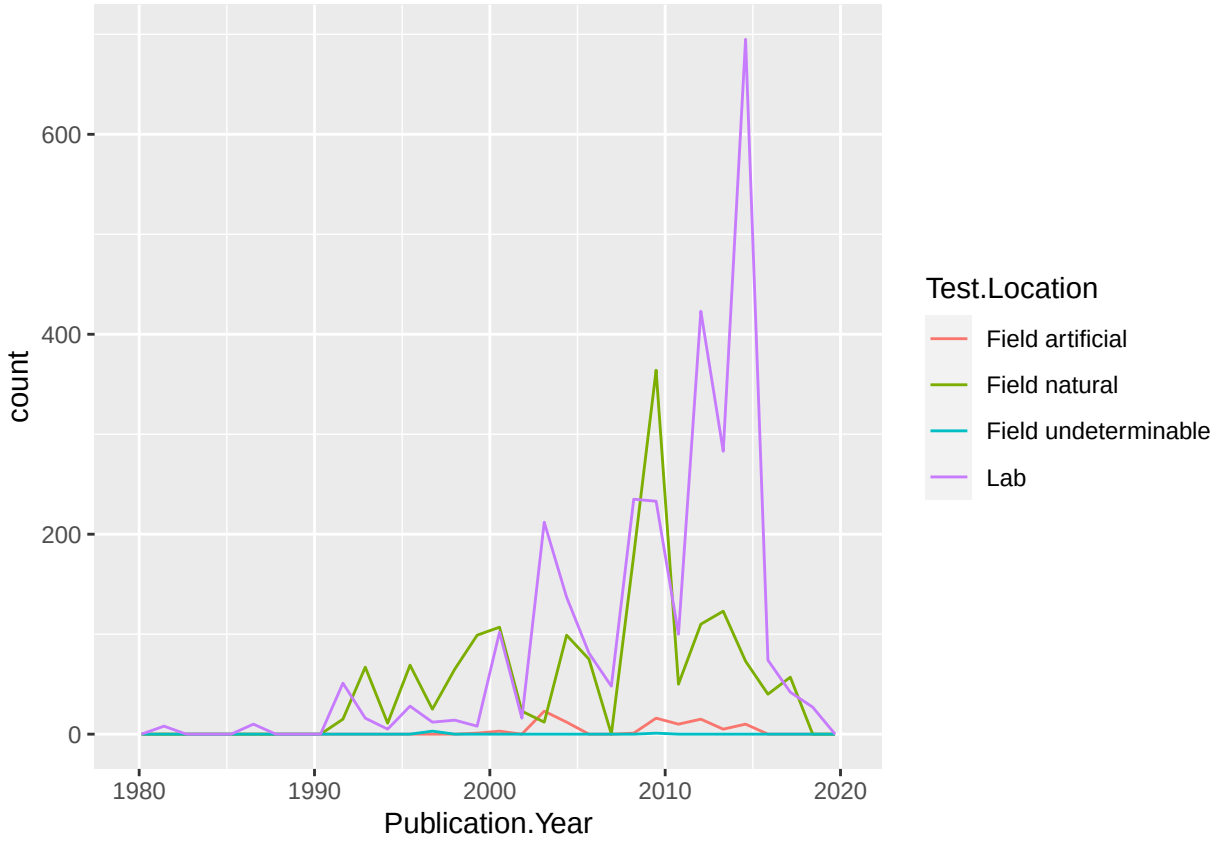
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

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

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

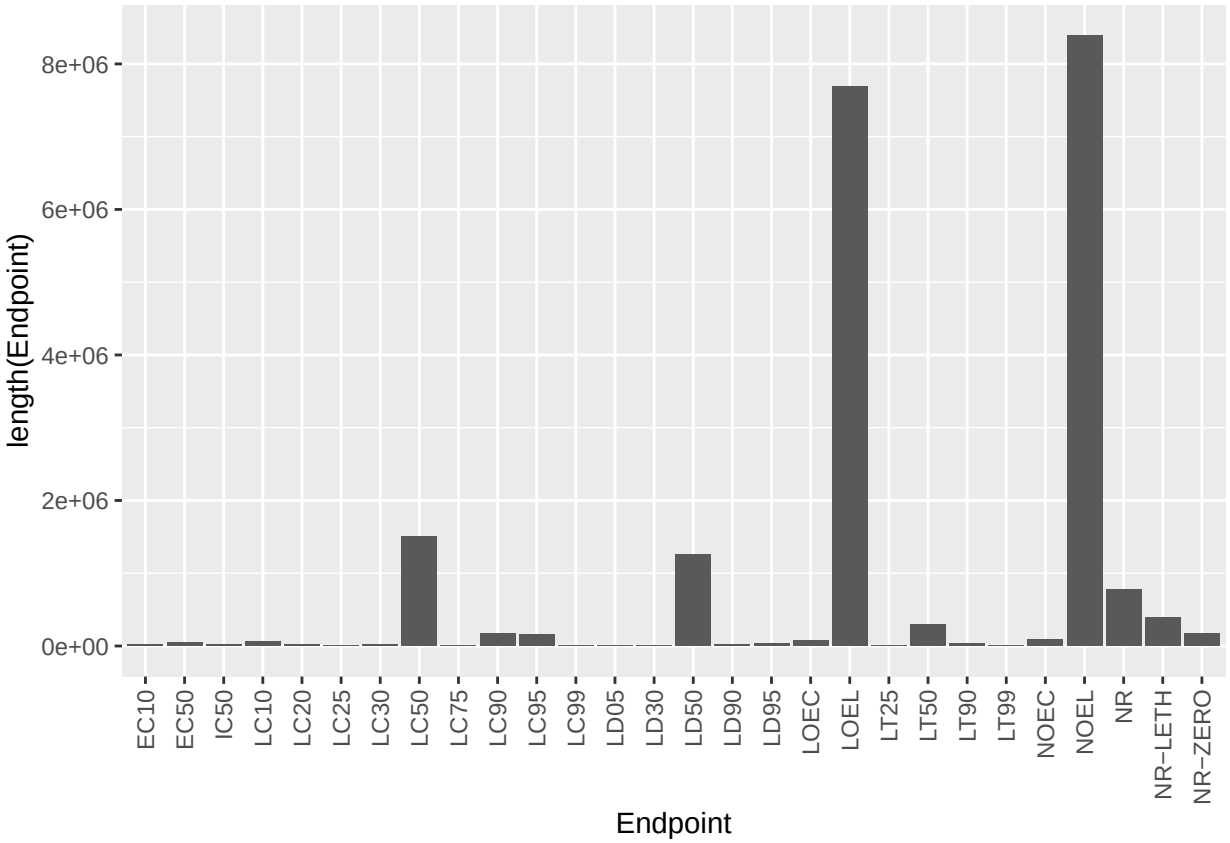


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: Lab and natural field test locations are the most common. Lab tests have consistently increased over time. Natural field tests were used pretty steadily and peaked in 2010 but have been decreasing in use since then.

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, y = length(Endpoint))) + geom_bar(stat = "identity") +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



Answer: The LOEL and NOEL are the two most common endpoints. LOEL is the lowest observable effect level and is defined by the lowest concentration of a substance that causes an observable effect when compared to controls. NOEL is the no observable effect level which is defined as the highest concentration of a substance which has no observable effect when compared to 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.

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

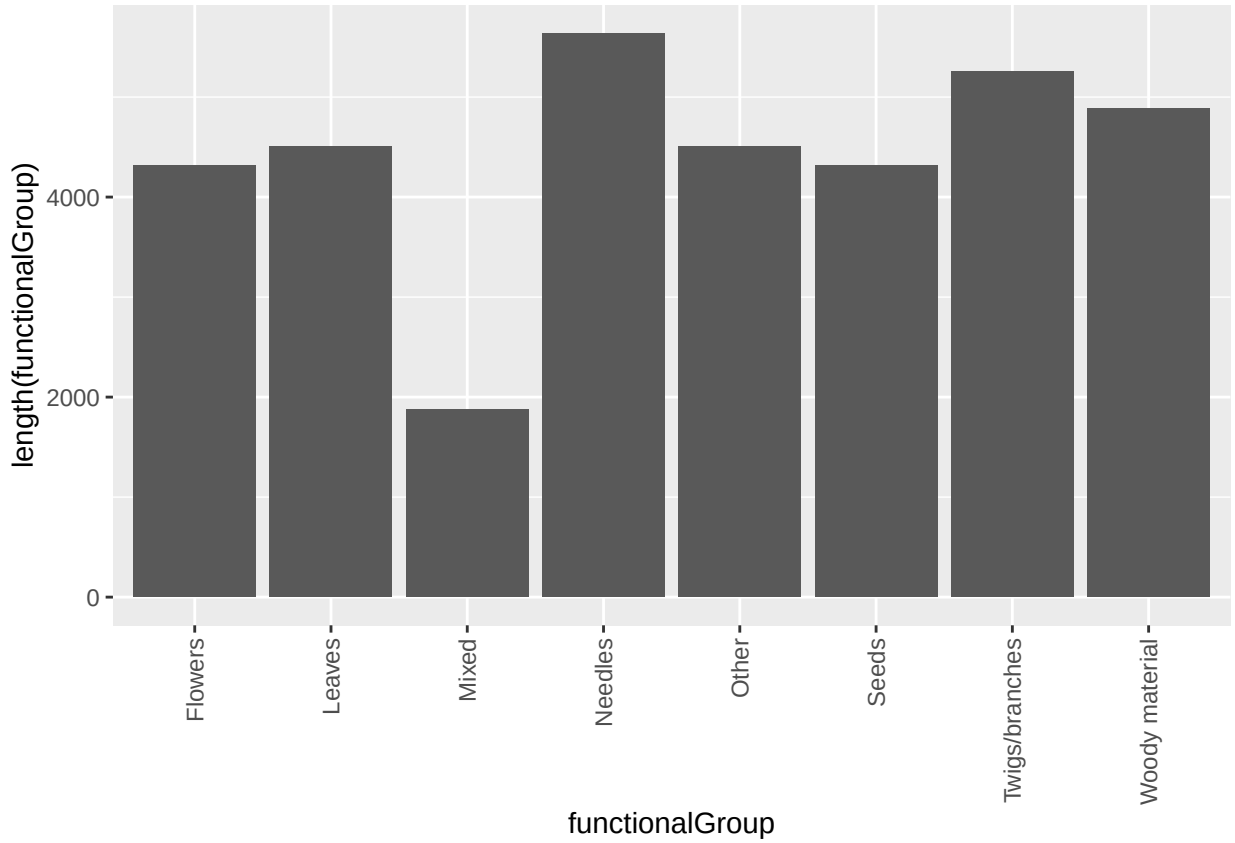
```
length(unique(Litter$plotID))
```

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

Answer: Duplicate values are removed using the unique function. Summary will report the number of duplicate values along with the values themselves.

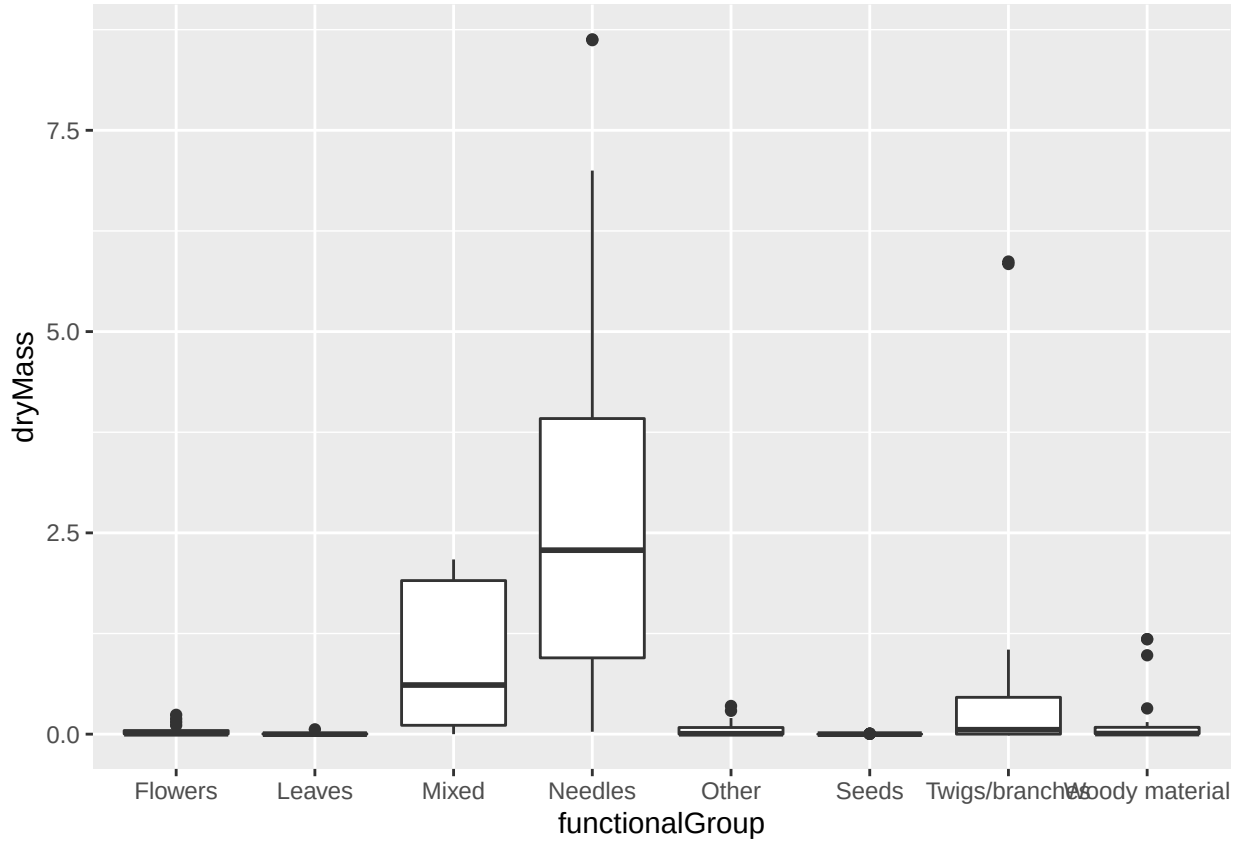
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

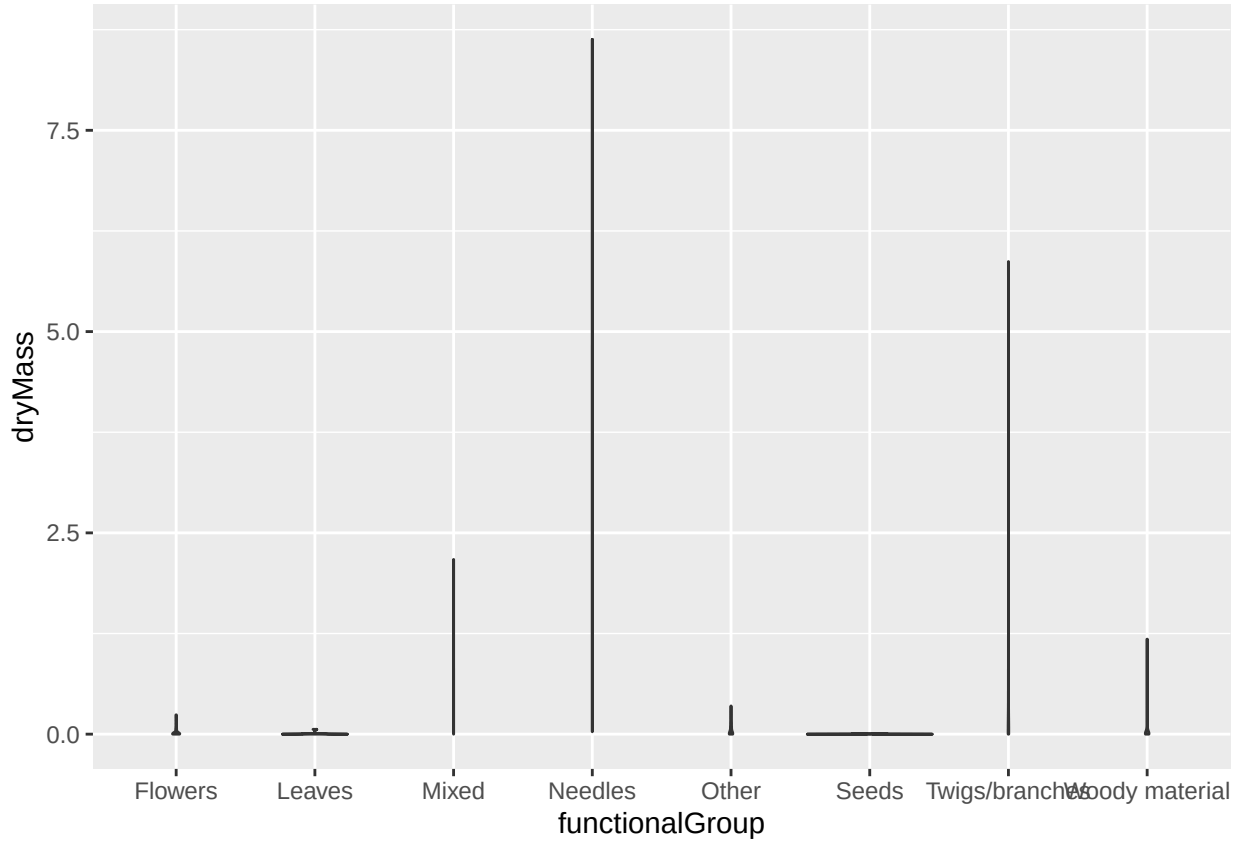


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
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: With this dataset the violin plot shows straight lines which do not tell us much about the summary statistics such as median or interquartile range. These items are more clearly communicated on the boxplot.

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

> Answer: Needles have the highest biomass followed by mixed and Twigs/branches/woody groups.