Assignment 3: Data Exploration

Sayra Martinez

Spring 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 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() #got my working directory
```

```
## [1] "/home/guest/EDA_Spring2024"
```

```
#Loaded the corresponding packages
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(lubridate)

#Reading the datasets
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",stringsAsFactors = T)
Litter <- read.csv("./Data/Raw/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: Because they migh have devastating ecological impacts and pose risk on agriculture production. First, because they affect both target (e.g., corn rootworm, flea beetle) and nontarget insects (e.g., bees), causing -when not death- chronic sublethal effects on aquatic insects, birds and pollinators, impacting agriculture. Second, due to they are highly soluble in water, easing their transport away from the area of initial application, making the problem easily widespread.

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: In the case of the woody debris, because of their ecological functions, being habitat for organisms, serve as an nutrient source,are a site for nitrogen fixation, and influence soil and sediment transport and storage. Besides,they play an essential role to the freshwater and estuarine ecosystems (the HJ Andrews Experimental Forest, 1986). Regarding the litter, this is directly involved in plant-soil interaction, helping to incorporate carbon and nutrients from plants into the soil, so it's relevant for nutrient cycling and, hence, in productivity in forest ecosystems (Giweta, 2020).

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: elevated and ground traps for litter. 2. Randomized trap placement in sites with >50% aerial cover of woody vegetation >2m in height.Targeted trap placement in sites with < 50% cover of woody vegetation. 3. Temporal sampling: ground traps are sampled once per year; while for elevated ones, varies according to vegetation present at the site (more frequent in deciduous forest sites during senescence, and infrequent year-round at evergreen sites).

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) #Consulting the dimension of my DS, I got that it has 4623 rows and 30 columns of data.
```

```
## [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?

```
# Getting the data only for Effect
sort(summary(Neonics$Effect), decreasing = TRUE)
```

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

Answer: The 3 most common studied effects are Population, Mortality and Behavior. With such information, we can try to test the Neonicotinoids' impacts on different species, not only changes in the population by higher levels of mortality, but also in behaviors, and try to test if they are attributable to neonicotinoids or not.

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), decreasing = T) #query for frequency by species (sorted from
```

```
##                                (Other)                              Honey Bee
##                                    670                                    667
##                         Parasitic Wasp                  Buff Tailed Bumblebee
##                                    285                                    183
##                    Carniolan Honey Bee                             Bumble Bee
##                                    152                                    140
##                       Italian Honeybee                        Japanese Beetle
##                                    113                                     94
```

```
##                 Asian Lady Beetle                Euonymus Scale
##                                76                            75
##                          Wireworm             European Dark Bee
##                                69                            66
##                 Minute Pirate Bug           Asian Citrus Psyllid
##                                62                            60
##                     Parastic Wasp        Colorado Potato Beetle
##                                58                            57
##                   Parasitoid Wasp           Erythrina Gall Wasp
##                                51                            49
##                      Beetle Order   Snout Beetle Family, Weevil
##                                47                            47
##          Sevenspotted Lady Beetle                True Bug Order
##                                46                            45
##              Buff-tailed Bumblebee                  Aphid Family
##                                39                            38
##                    Cabbage Looper         Sweetpotato Whitefly
##                                38                            37
##                     Braconid Wasp                  Cotton Aphid
##                                33                            33
##                    Predatory Mite        Ladybird Beetle Family
##                                33                            30
##                        Parasitoid                 Scarab Beetle
##                                30                            29
##                     Spring Tiphia                   Thrip Order
##                                29                            29
##              Ground Beetle Family            Rove Beetle Family
##                                27                            27
##                     Tobacco Aphid                  Chalcid Wasp
##                                27                            25
##           Convergent Lady Beetle                 Stingless Bee
##                                25                            25
##                 Spider/Mite Class           Tobacco Flea Beetle
##                                24                            24
##                  Citrus Leafminer               Ladybird Beetle
##                                23                            23
##                         Mason Bee                      Mosquito
##                                22                            22
##                    Argentine Ant                         Beetle
##                                21                            21
##       Flatheaded Appletree Borer           Horned Oak Gall Wasp
##                                20                            20
##                Leaf Beetle Family             Potato Leafhopper
##                                20                            20
##       Tooth-necked Fungus Beetle                  Codling Moth
##                                20                            19
##        Black-spotted Lady Beetle                  Calico Scale
##                                18                            18
##              Fairyfly Parasitoid                   Lady Beetle
##                                18                            18
##           Minute Parasitic Wasps                     Mirid Bug
##                                18                            18
##                  Mulberry Pyralid                      Silkworm
##                                18                            18
```

```
##                    Vedalia Beetle             Araneoid Spider Order
##                                18                                17
##                         Bee Order                    Egg Parasitoid
##                                17                                17
##                      Insect Class        Moth And Butterfly Order
##                                17                                17
##      Oystershell Scale Parasitoid Hemlock Woolly Adelgid Lady Beetle
##                                17                                16
##             Hemlock Wooly Adelgid                              Mite
##                                16                                16
##                       Onion Thrip             Western Flower Thrips
##                                16                                15
##                      Corn Earworm                Green Peach Aphid
##                                14                                14
##                         House Fly                        Ox Beetle
##                                14                                14
##                 Red Scale Parasite                Spined Soldier Bug
##                                14                                14
##             Armoured Scale Family                  Diamondback Moth
##                                13                                13
##                      Eulophid Wasp                Monarch Butterfly
##                                13                                13
##                      Predatory Bug             Yellow Fever Mosquito
##                                13                                13
##                Braconid Parasitoid                      Common Thrip
##                                12                                12
##      Eastern Subterranean Termite                            Jassid
##                                12                                12
##                        Mite Order                         Pea Aphid
##                                12                                12
##                   Pond Wolf Spider         Spotless Ladybird Beetle
##                                12                                11
##             Glasshouse Potato Wasp                          Lacewing
##                                10                                10
##            Southern House Mosquito           Two Spotted Lady Beetle
##                                10                                10
##                        Ant Family                      Apple Maggot
##                                 9                                 9
```

Answer:The most common estudied species are the Honey Bee, the Parasitic Wasp, the Buff Tailed Bumblebee, the Carniolan Honey Bee, the Bumble Bee, and the Italian Honeybee.

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"
```

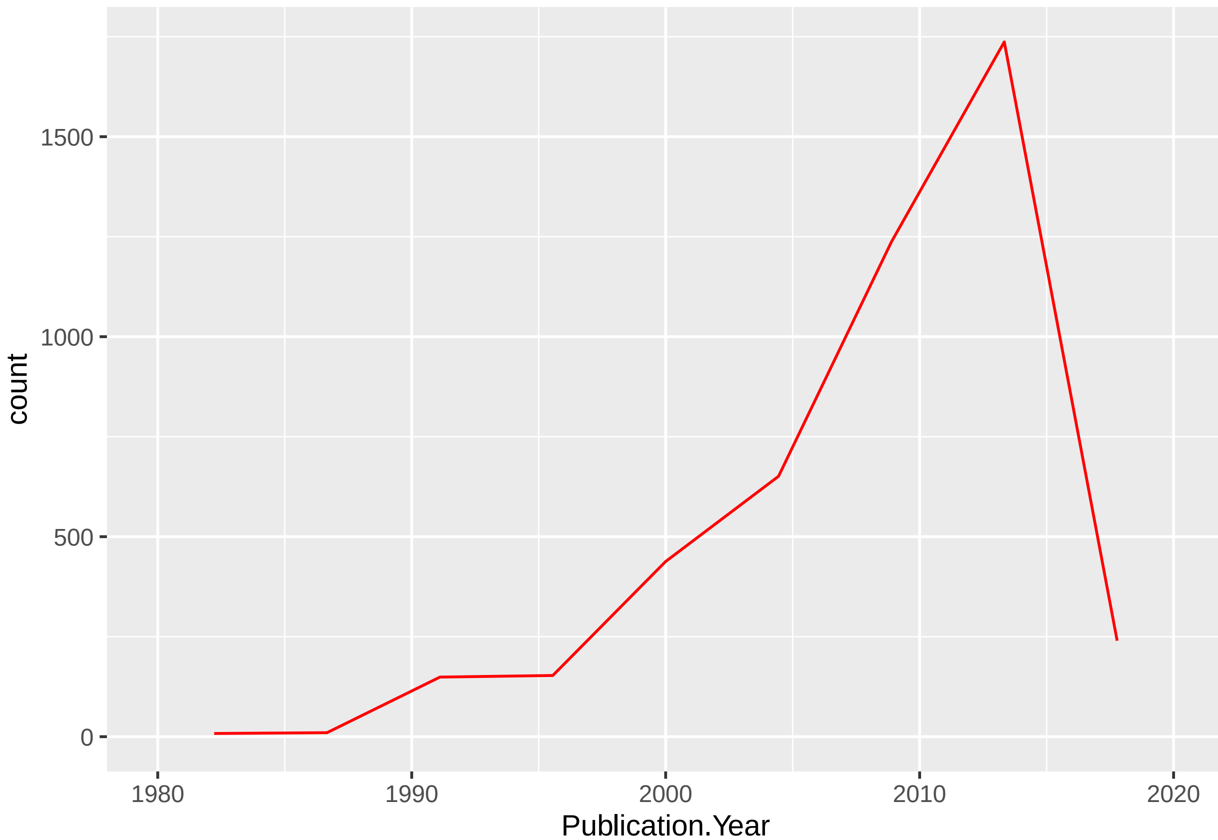
Answer: It is a Factor type, as contains symbols such as "~" and "/".

Explore your data graphically (Neonics)

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

```
#Create the plot
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 10, color = "red",  lty = 1) +  scale_x_continuous(lin
```

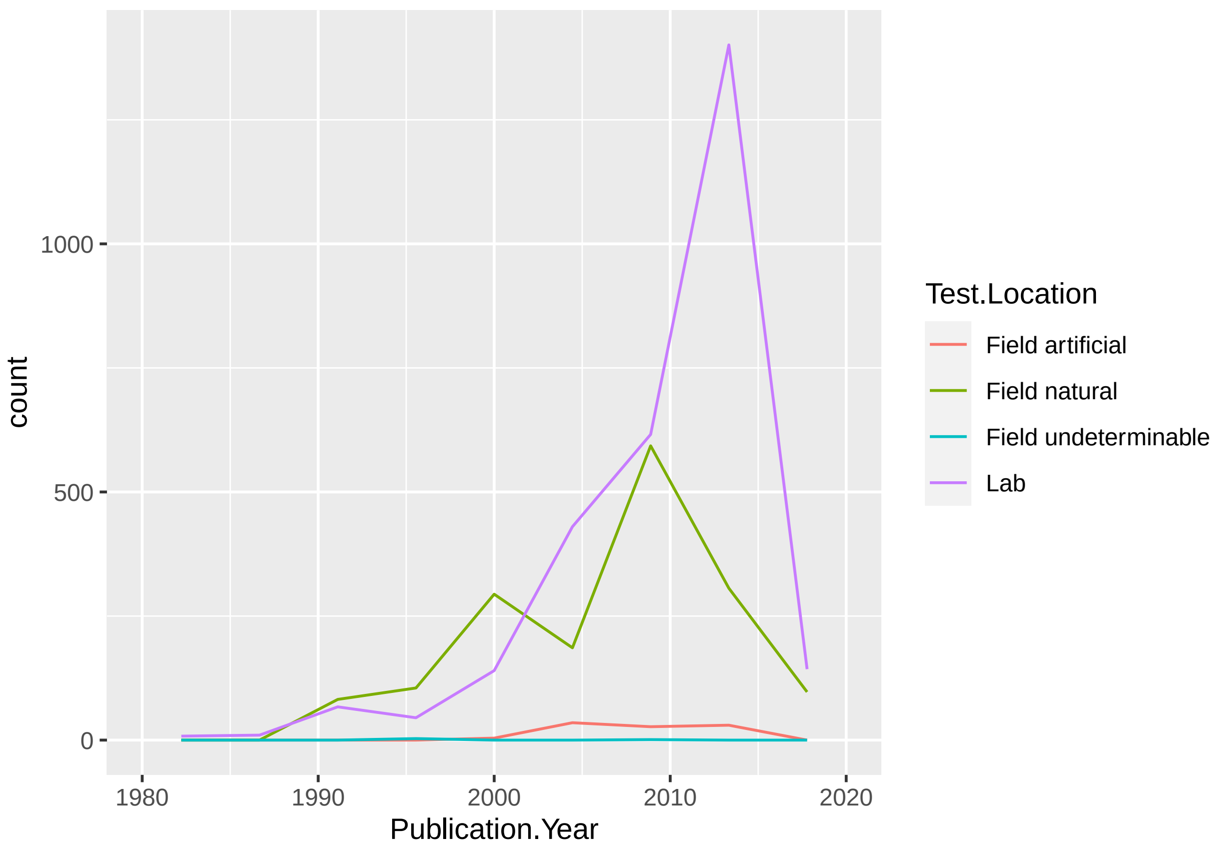
```
## Warning: Removed 2 rows containing missing values (`geom_path()`).
```



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, group = Test.Location, color = Test.Location), bins = 10, lty
  scale_x_continuous(limits = c(1980, 2020))
```

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



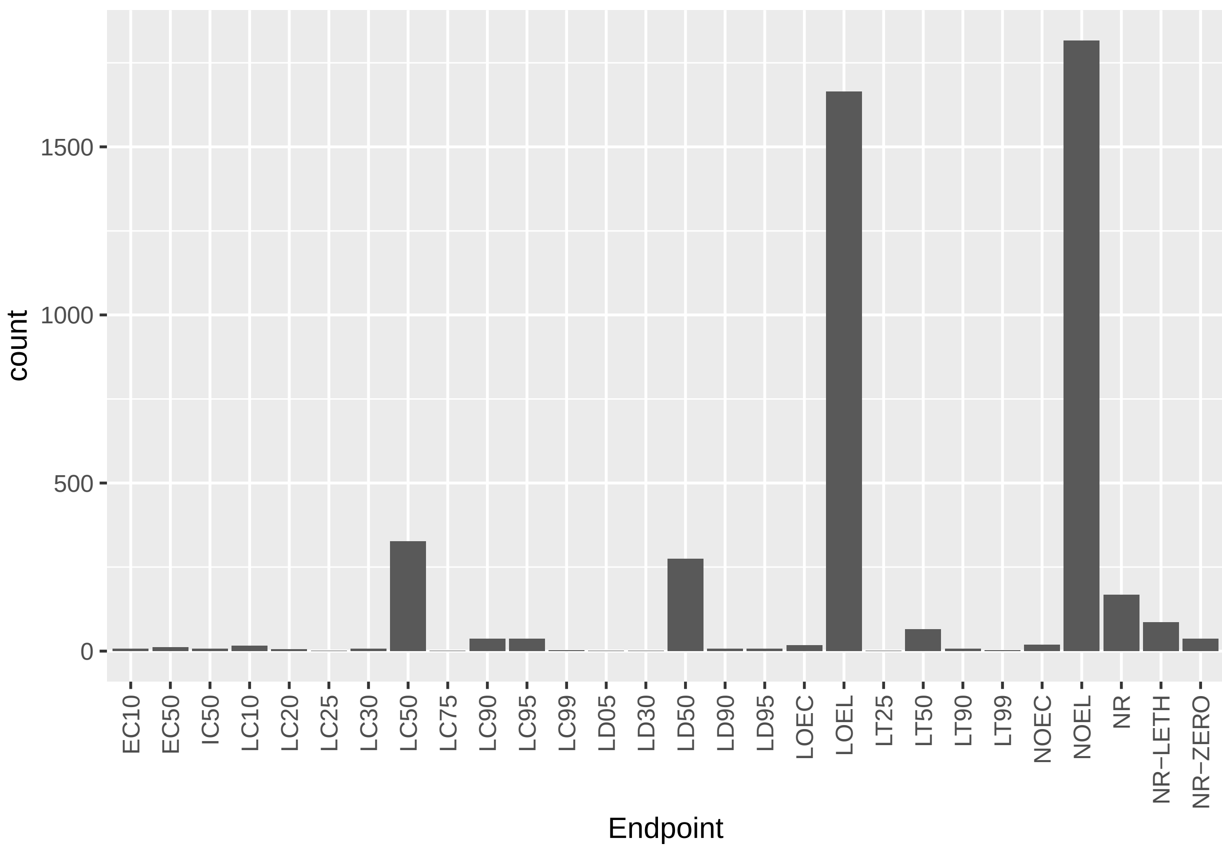
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: All along the plot, we observe that Lab and Field natural are the most common test locations. Although "Field natural" were the main type of test location for decades, in early 2000's years the use of lab gain relevance an became the most used type. Even for the last years, when the number of publications dropped dramatically, the lab is still more used than field natural.

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



Answer: The two most common are: 1) No-observable-effect-level (NOEL), which means that the highest dose (concentration) does not produce effects significantly different from responses of controls; and 2) Lowest-observable-effect-level (LOEL), that means that lowest dose (concentration) produce effects 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.

```
class(Litter$collectDate) #It's a factor type, so below I will put it as a date type
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y/%m/%d")
class(Litter$collectDate) #Now it is Date type
```

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

```
Augustsampled_litter <- unique(Litter$collectDate[format(Litter$collectDate, "%Y/%m")  == 2018/08])
print(Augustsampled_litter) #None of the values were unique, so all of them were sampled in August 2018
```

```
## [1] NA
```

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`?

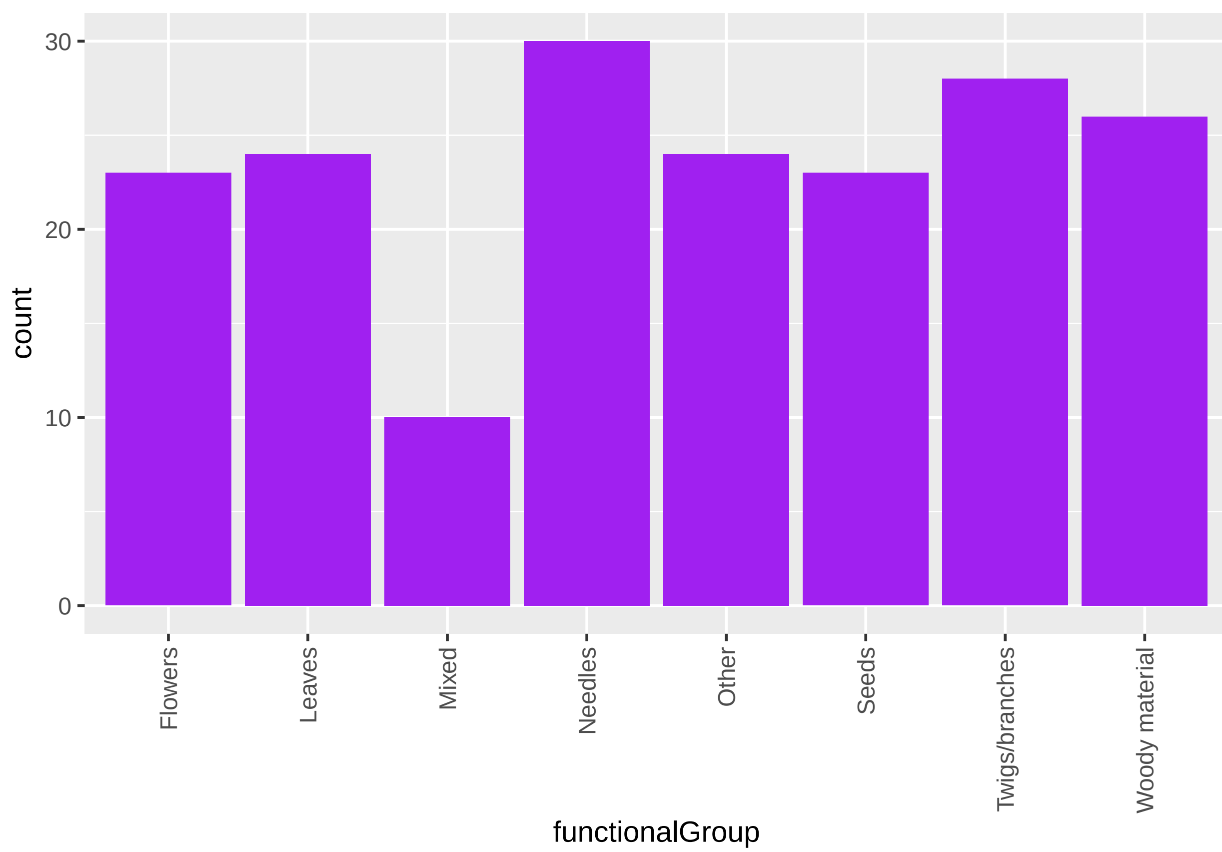
```
Niwot_Ridged_sampled <- unique(Litter$plotID)
print(Niwot_Ridged_sampled) #12 Plots were sampled
```

```
##  [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: unique() function eliminates the duplicate values or the rows, and returns only those that are unique; while summary returns the frequency of each value.

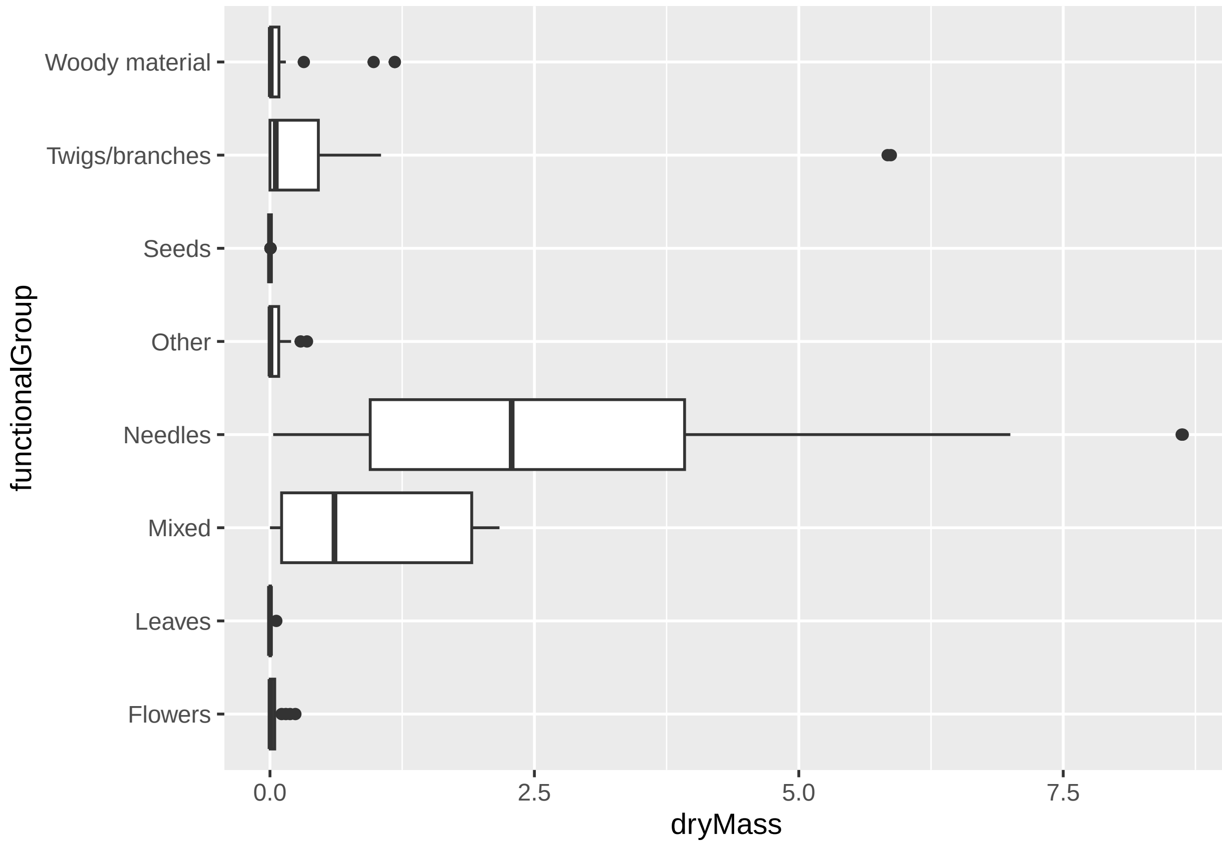
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

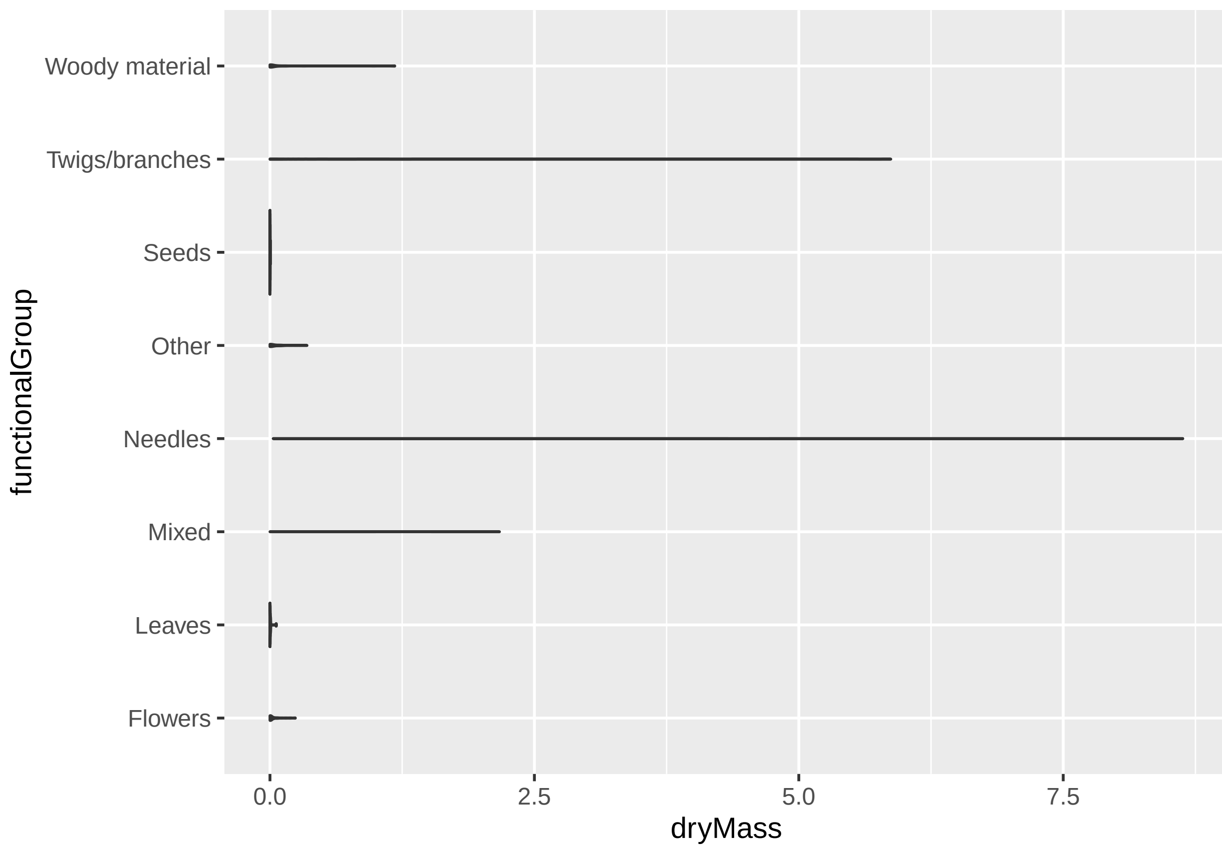


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




```r
#For geom_violin
ggplot(Litter) +
  geom_violin(aes(x = dryMass, y = functionalGroup),
              draw_quantiles = c(0.25, 0.5, 0.75))
```



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

> Answer: Because the violin plot shows distribution and as the size of the sample is small with several functional groups, it's not the most suitable option.

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

> Answer: The needles.