Assignment 3: Data Exploration

Gary Alvarez

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() #I checked that my Working directory is set to the base folder for
```

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

```
#the Environmental Data Analytics Course repository.
#install.packages("lubridate"), tidyverse and here.
library(lubridate)
library(tidyverse)
library(here)
library(ggplot2)
```

```
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
  )
#I uploaded the two datasets.
```

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:Studying the ecotoxicology of neonicotinoids on insects is essential for balancing the benefits of pest control in agriculture with the potential risks to biodiversity, ecosystem functioning, and human well-being. It helps guide sustainable agricultural practices and the development of alternative pest management strategies.

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: Studying litter and woody debris in forest ecosystems provides a comprehensive understanding of the ecological processes that shape these environments. This knowledge is critical for sustainable forest management, conservation efforts, and predicting the impacts of environmental changes on ecosystems.

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 and fine woody debris are collected from elevated and ground traps, respectively. 2. Ground traps are sampled once per year. Target sampling frequency for elevated traps varies by vegetation present at the site, with frequent sampling (1x every 2weeks) in deciduous forest sites during senescence, and in-frequent year-round sampling (1x every 1-2 months) at evergreen sites. 3. In this protocol, litter is defined as material that is dropped from the forest canopy and has a butt end diameter <2cm and a length <50 cm; this material is collected in elevated 0.5m2 PVC traps. Fine wood debris is defined as material that is dropped from the forest canopy and has a butt end diameter <2cm and a length >50 cm; this material is collected in ground traps as longer material is not reliably collected by the elevated traps.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
class(Neonics) #The type of file is data frame.
```

```
## [1] "data.frame"
```

```
colnames(Neonics) #We can see the names of the 30 columns that make up the data frame.
```

```
##  [1] "CAS.Number"                      "Chemical.Name"
##  [3] "Chemical.Grade"                  "Chemical.Analysis.Method"
##  [5] "Chemical.Purity"                 "Species.Scientific.Name"
##  [7] "Species.Common.Name"             "Species.Group"
##  [9] "Organism.Lifestage"              "Organism.Age"
## [11] "Organism.Age.Units"              "Exposure.Type"
## [13] "Media.Type"                      "Test.Location"
## [15] "Number.of.Doses"                 "Conc.1.Type..Author."
## [17] "Conc.1..Author."                 "Conc.1.Units..Author."
## [19] "Effect"                          "Effect.Measurement"
## [21] "Endpoint"                        "Response.Site"
## [23] "Observed.Duration..Days."        "Observed.Duration.Units..Days."
## [25] "Author"                          "Reference.Number"
## [27] "Title"                           "Source"
## [29] "Publication.Year"                "Summary.of.Additional.Parameters"
```

```
dim(Neonics) #Neonics 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:The three most common effects studied are Population, Mortality, and Behavior. These effects might specifically be of interest because they offer information about the population size of insects at a given time (Population), the rate at which they are dying (Mortality), and the general behavior they express (Behavior).

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...]

```
sum_common_names <- summary(Neonics$Species.Common.Name)
sort(sum_common_names, decreasing = TRUE) #Sorting the output of the summary line
```

```
##                          (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: Excluding the category Other (670), the six most commonly studied species in the dataset are: 1. Honey Bee (667), 2. Parasitic Wasp (285), 3. Buff Tailed Bumblebee (183), 4. Carniolan Honey Bee (152), 5. Bumble Bee (140), 6. Italian Honeybee (113). All these species are pollinators, and they are more important over other species because the insecticides kill all insects within a range of land. That includes pollinators, whose work is benefitial to the 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?

```
class(Neonics$Conc.1..Author)
```

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

```
summary(Neonics$Conc.1..Author)
```

```
##    0.37/      10/      NR/       NR        1     1023    0.40/       2/
##      208      127      108       94       82       80       69       63
##       10   0.053/      100      50/     0.5/     0.03    0.05/     0.45
##       62       59       56       51       45       44       43       43
##     0.1/    0.45/     1.0/    2.27/       50    0.125     500/      0.5
##       42       40       40       40       36       33       33       32
##   0.048/    0.15/       1/       48    25.0/       12/    0.027      2.4
##       30       30       30       30       28       27       26       26
##     0.2/    0.56/     100/        3    0.01/     1000/       3/    0.336
##       25       24       23       23       22       22       22       21
##     1.5/     0.05      1.5    2.60/    20.0/        6    6.80/    62.5/
##       21       20       20       20       20       20       20       20
##    0.005     0.4/    0.18/     0.3/     1000       40 0.00355/      0.1
##       18       18       17       17       17       17       16       16
##      0.4     150/      300      80/    0.053     0.24     0.28     125/
##       16       16       16       16       15       15       15       15
##        9   0.0001  0.0004/   0.084/     0.15      0.6    12.5/   144.0/
##       15       14       14       14       14       14       14       14
##     350/    40.0/      48/       56      84/    0.17/      125       14
##       14       14       14       14       14       13       13       13
##       16       17   0.047/    0.25/    0.28/    1.28/    1.81/      112
##       13       13       12       12       12       12       12       12
##      150     2.5/       25      60/      75/    0.02/   0.025/     0.29
##       12       12       12       12       12       11       11       11
##    37.5/       4/        5  (Other)
##       11       11       11     1817
```

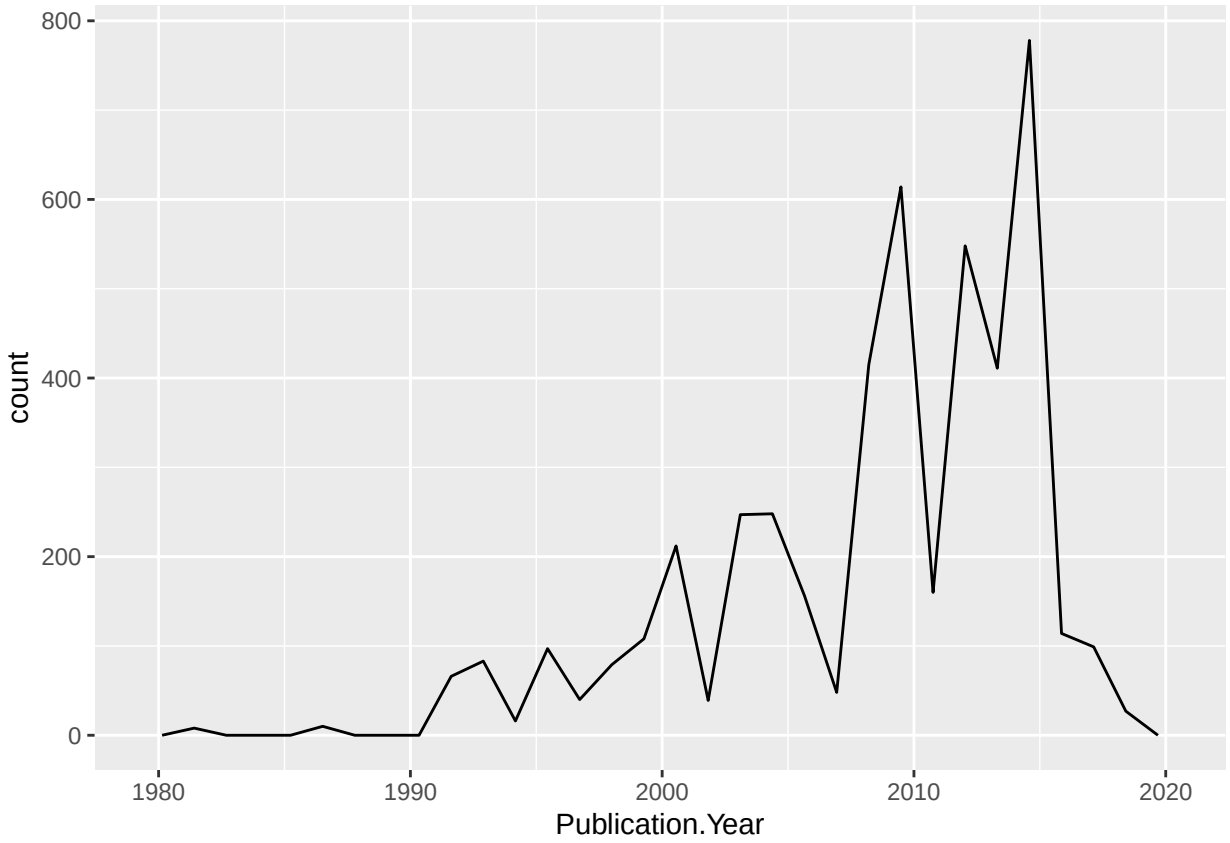
Answer: The class of `Conc.1..Author.` is factor, because it represents categories.

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))
```

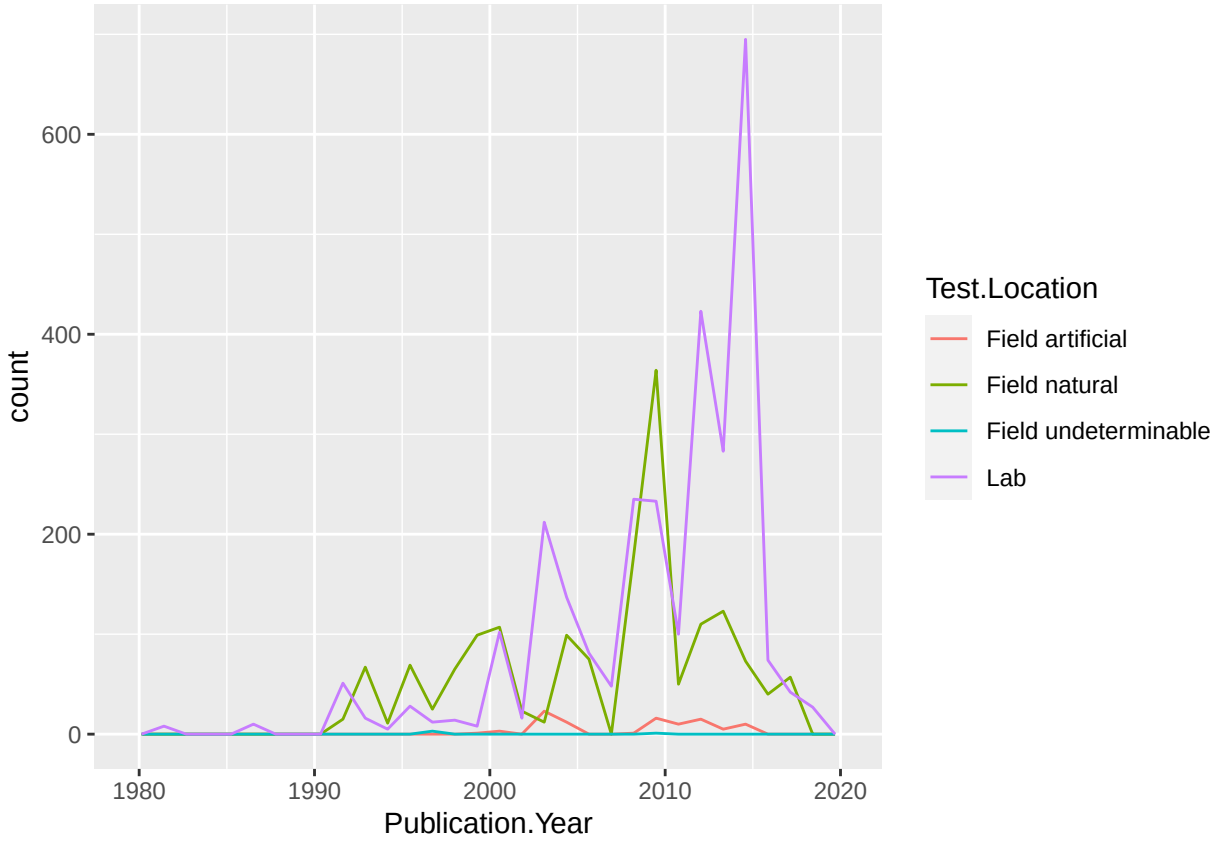
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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`.
```



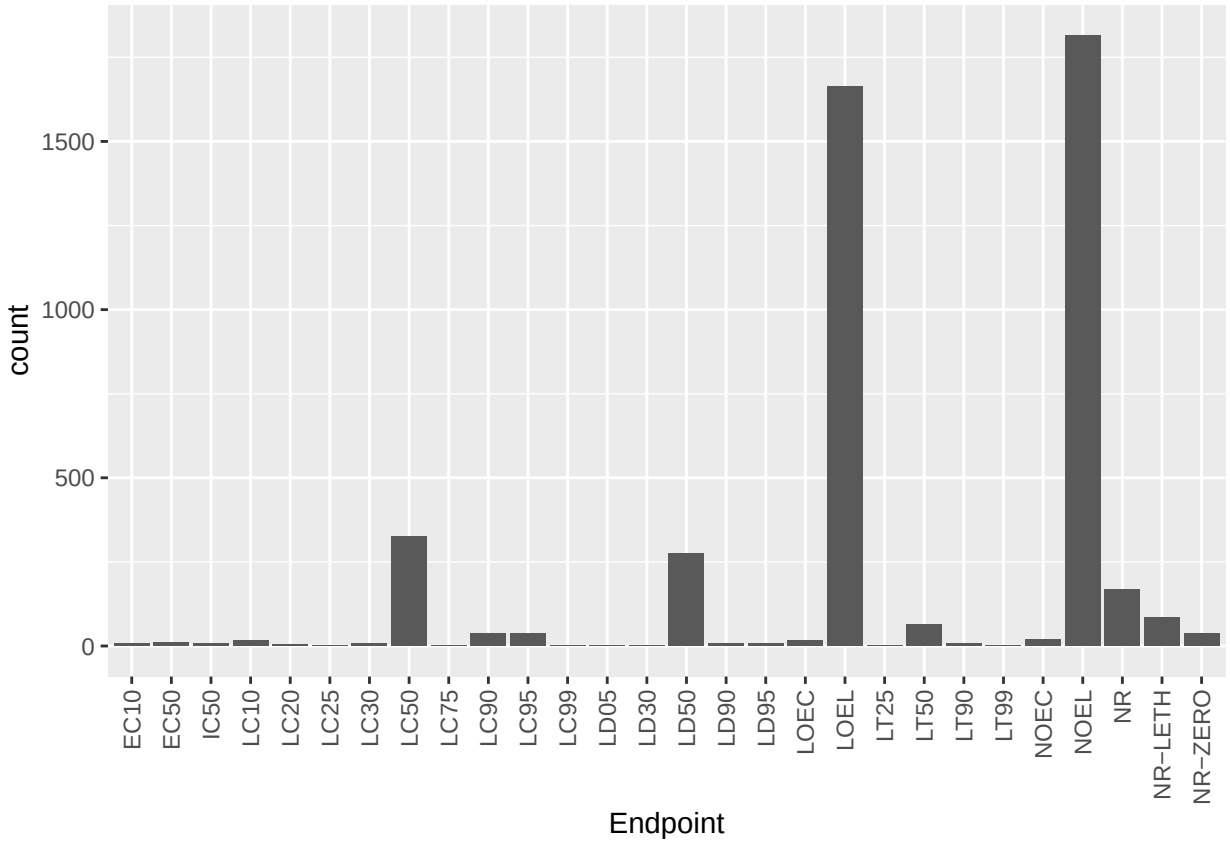
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are Lab and Fielt natural. They change over time. From 1990 to 2000, Field artificial and Lab were the most common locations. However, after a peak of Field natural in 2009, this location became significantly less common than Lab, and this trends holds until 2020.

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))
```



Answer: The most common endpoints are NOEL and LOEL. The terms "NOEL" (No Observed Effect Level) and "LOEL" (Lowest Observed Effect Level) are used in toxicology and risk assessment to describe exposure levels at which no adverse effects are observed (NOEL) or the lowest exposure level at which adverse effects are observed (LOEL).

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) #The class of collectDate is originally "factor".
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
unique(Litter$collectDate, incomparables = FALSE)
```

```
## [1] "2018-08-02" "2018-08-30"
```

Answer: The dates litter was sampled in August 2018 are August 02 and August 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, incomparables = FALSE)
```

```
##  [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
```

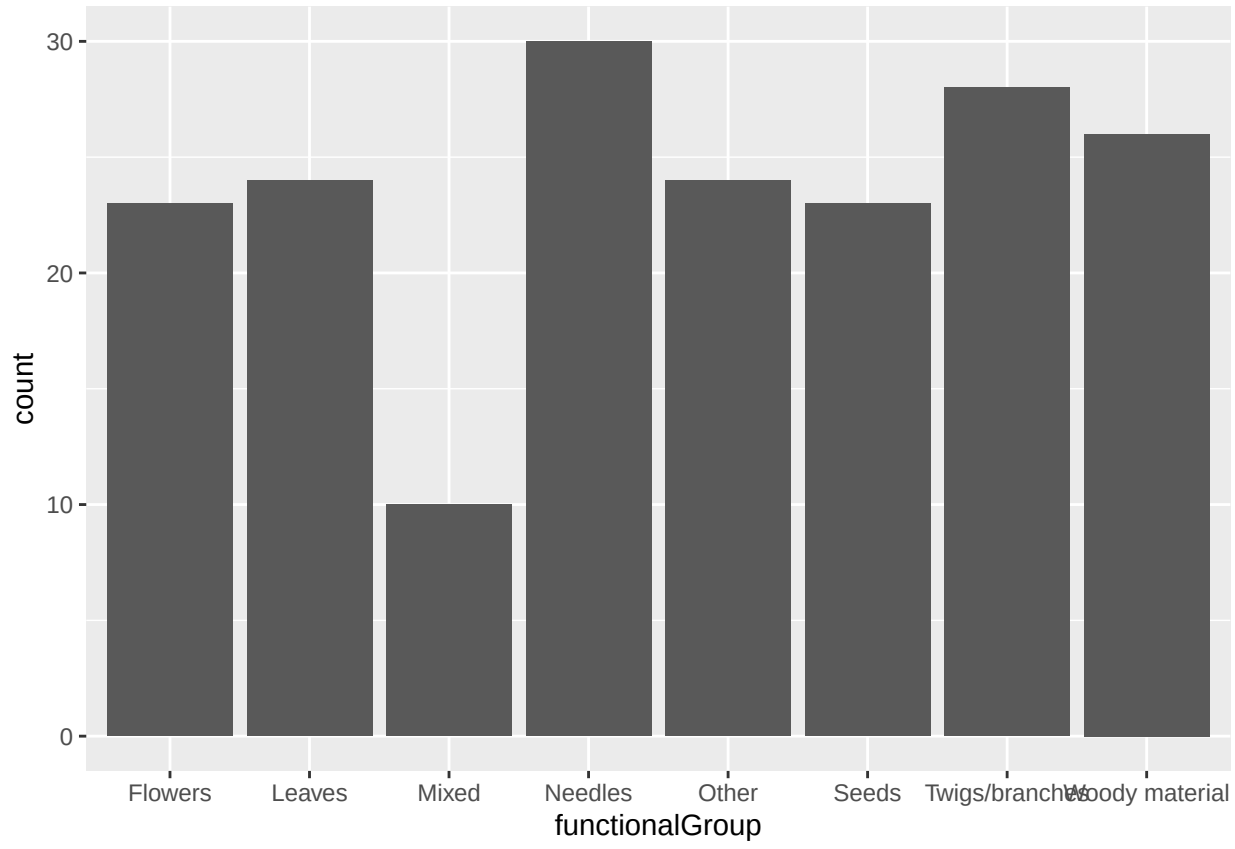
```
summary(Litter$plotID)
```

```
## 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
```

Answer: 12 plots were sampled at Niwot Ridge. The 'unique' function is used to obtain the unique values of a vector or a column in a data frame. It returns a vector containing the unique elements in the order in which they appear in the input. On the other hand, the 'summary' function provides a summary of the distribution of a variable, typically for numerical variables. Since plotID is not a numeric variable, it is better to use the 'unique' function.

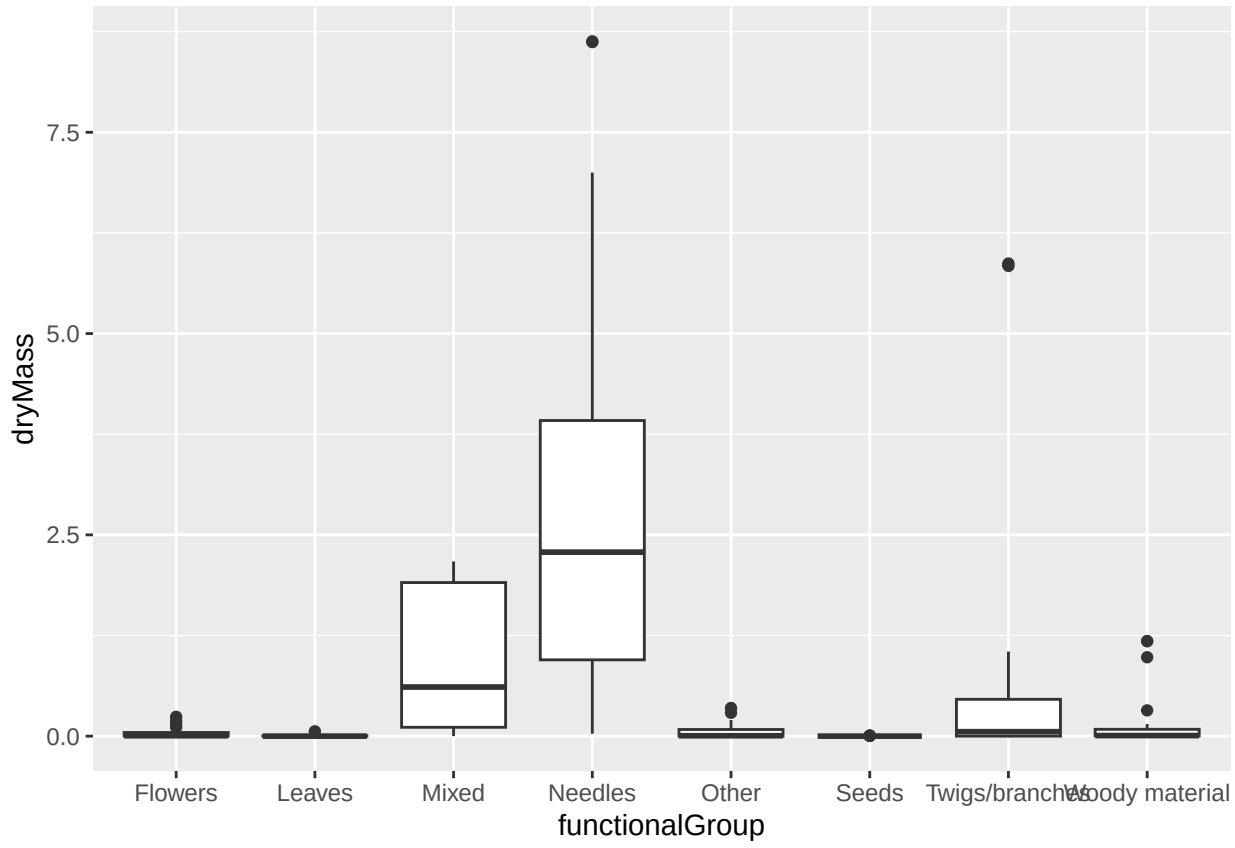
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(x =functionalGroup))+
       geom_bar()
```

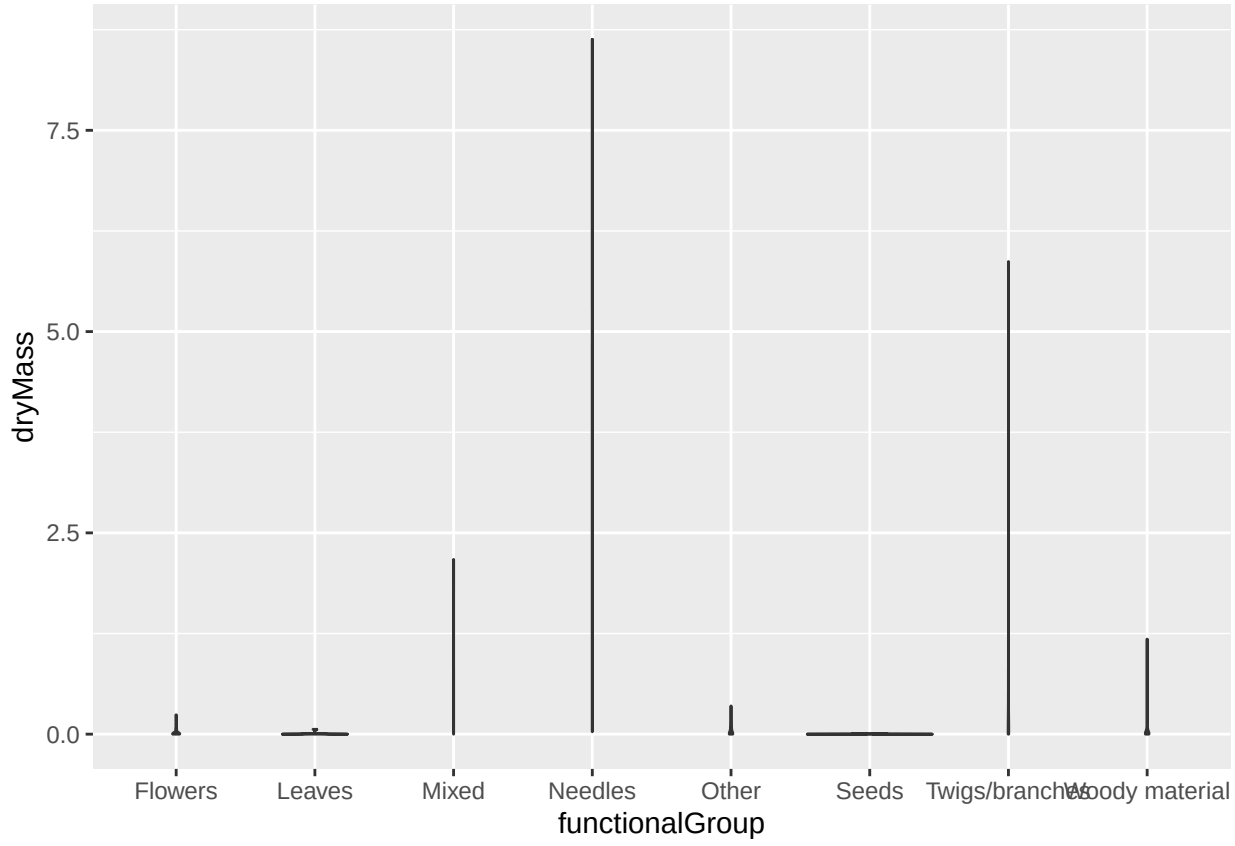


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




```
ggplot(Litter)+
       geom_violin(aes(x =functionalGroup, y=dryMass))
```



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

    Answer: Because the boxplot is able to convey more meaningful information about the distribution. With the characteristics of this dataset, the violin plot looks like multiple vertical lines, and that doesn't provide much information.

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

Answer: Needles, and Twigs/branches.