Assignment 3: Data Exploration

Susanna Jenkins

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()
```

```
## [1] "/home/guest/EDA-Spring2023"
```

```
library(tidyverse)
library(lubridate)

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: The ecotoxicology of neonicotinoids on insects is important to study because it poses a threat to pollinators, and could potentially be causing the decline in honey bees. As environmental graduate students, it is important that we understand this impact to polliantors, given their importance.

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: It is important to study litter and woody debris that falls to the forest ground because woody debris may act as a tinder that promotes the start and spreading of forest fires, and forest fires are becoming increasingly frequent and intense with our warming climate.

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. The literfall and fine woody debris sampling data products provide mass data for plant functional groups from individual sampling bouts. Liter and fine woody debris are collected from elevated and ground traps. 2.Liter and fine woody debris sampling is executed at terrestrial NEON sites that contain woody vegetation >2m tall. 3. Along with most of NEON's plant productivity measurements, sampling for this product occurs only in towerplots. Locations of tower plots are selected randomly within the 90% flux footprint of the primary and secondary airsheds.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
str(Neonics) # 4623 obs. of  30 variables
```

```
## 'data.frame':    4623 obs. of  30 variables:
##  $ CAS.Number                      : int  58842209 58842209 58842209 58842209 58842209 58842209 5884
##  $ Chemical.Name                   : Factor w/ 9 levels "(1E)-N-[(6-Chloro-3-pyridinyl)methyl]-N-ethy
##  $ Chemical.Grade                  : Factor w/ 9 levels "Analytical grade",..: 9 9 9 9 9 9 9 9 9 9 .
##  $ Chemical.Analysis.Method        : Factor w/ 5 levels "Measured","Not coded",..: 4 4 4 4 4 4 4 4 4
##  $ Chemical.Purity                 : Factor w/ 80 levels ">=98",">=99.0",..: 69 69 50 50 50 50 50 50
##  $ Species.Scientific.Name         : Factor w/ 398 levels "Acalolepta vastator",..: 69 69 248 248 248
##  $ Species.Common.Name             : Factor w/ 303 levels "Alfalfa Leafcutter Bee",..: 74 74 142 142
##  $ Species.Group                   : Factor w/ 4 levels "Insects/Spiders",..: 1 1 1 1 1 1 1 1 1 1 ..
##  $ Organism.Lifestage              : Factor w/ 20 levels "Adult","Cocoon",..: 1 1 19 19 19 1 19 1 1 1
```

```
##  $ Organism.Age                    : Factor w/ 39 levels "<=24","<=48",..: 39 39 39 39 39 36 39 36 3
##  $ Organism.Age.Units              : Factor w/ 11 levels "Day(s)","Days post-emergence",..: 9 9 4 4 4
##  $ Exposure.Type                   : Factor w/ 24 levels "Choice","Dermal",..: 23 23 11 11 11 11 11 1
##  $ Media.Type                      : Factor w/ 10 levels "Agar","Artificial soil",..: 7 7 3 3 3 3 3 3
##  $ Test.Location                   : Factor w/ 4 levels "Field artificial",..: 4 4 4 4 4 4 4 4 4 4 .
##  $ Number.of.Doses                 : Factor w/ 30 levels "' 4-5","' 4-7",..: 30 30 18 18 18 18 18 18
##  $ Conc.1.Type..Author.            : Factor w/ 3 levels "Active ingredient",..: 1 1 1 1 1 1 1 1 1 1
##  $ Conc.1..Author.                 : Factor w/ 1006 levels "<0.0004","<0.025",..: 639 510 813 622 442
##  $ Conc.1.Units..Author.           : Factor w/ 148 levels "%","% v/v","% w/v",..: 132 132 91 91 91 91
##  $ Effect                          : Factor w/ 19 levels "Accumulation",..: 16 16 16 16 16 16 16 16 1
##  $ Effect.Measurement              : Factor w/ 155 levels "Abundance","Accuracy of learned task, peri
##  $ Endpoint                        : Factor w/ 28 levels "EC10","EC50",..: 15 15 8 8 8 8 8 8 8 8 ...
##  $ Response.Site                   : Factor w/ 19 levels "Abdomen","Brain",..: 14 14 14 14 14 14 14 1
##  $ Observed.Duration..Days.        : Factor w/ 361 levels "<.0002","<.0021",..: 145 145 145 145 145 1
##  $ Observed.Duration.Units..Days.  : Factor w/ 17 levels "Day(s)","Day(s) post-emergence",..: 1 1 1 1
##  $ Author                          : Factor w/ 433 levels "Abbott,V.A., J.L. Nadeau, H.A. Higo, and M
##  $ Reference.Number                : int  107388 107388 103312 103312 103312 103312 103312 103312 103
##  $ Title                           : Factor w/ 458 levels "A Common Pesticide Decreases Foraging Succ
##  $ Source                          : Factor w/ 456 levels "Acta Hortic.1094:451-456",..: 295 295 296
##  $ Publication.Year                : int  1982 1982 1986 1986 1986 1986 1986 1986 1986 1986 ...
##  $ Summary.of.Additional.Parameters: Factor w/ 943 levels "Purity: \xca NC - NC | Organism Age: \xca
```

```
dim(Neonics) # 4623 rows and 30 columns
```

```
## [1] 4623   30
```

```
length(Neonics) # 30 columns
```

```
## [1] 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: Most common effects studied include: Population, Mortality, Behavior, Feeding Behavior, Reproduction. These effects might be of specific interest as they are more easily observed and can tell a lot about a 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: 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 6 most commonly studied species are: Other, Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee. They might be of interest as they are pollinators and provide critical services to our earth's environmental functions, and are at risk.

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"
```

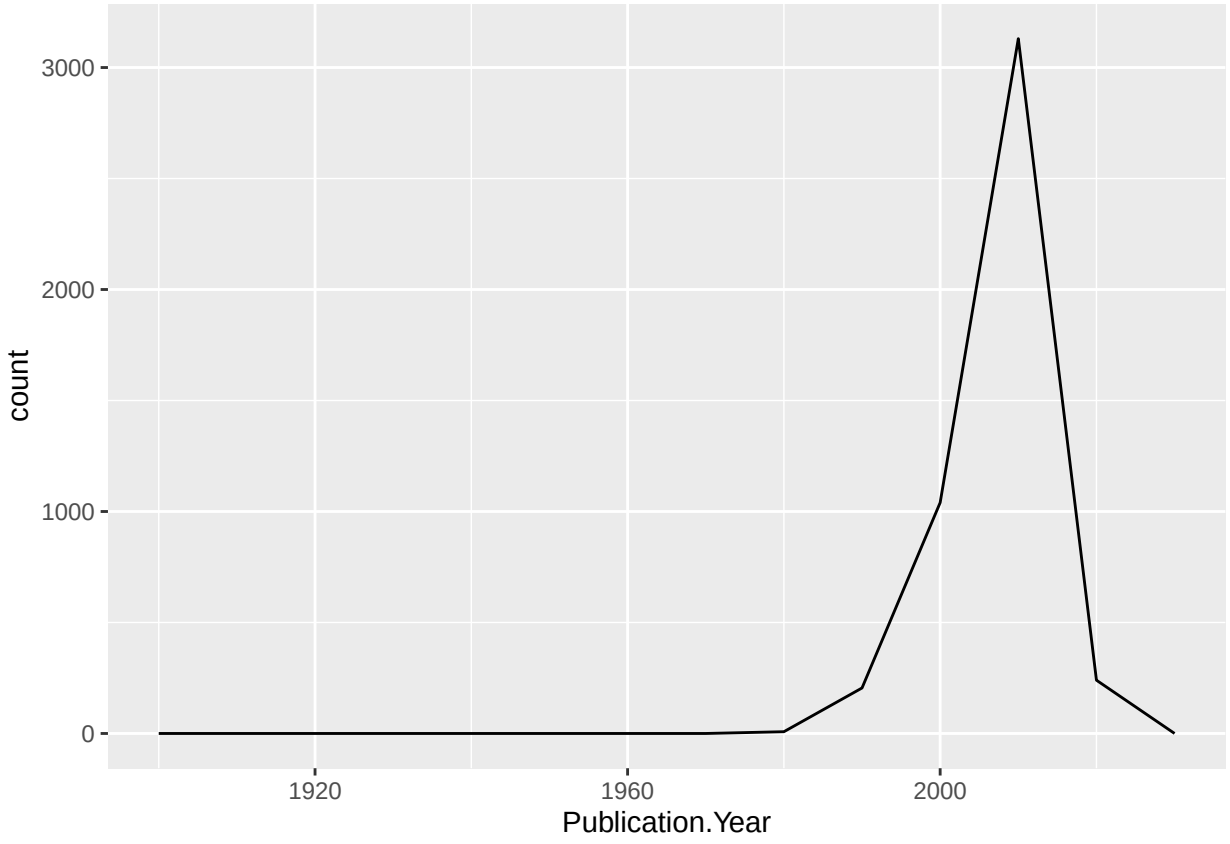
Answer: It is a factor; it is not numeric because it contains both words and numerical values.

Explore your data graphically (Neonics)

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

```
ggplot(Neonics, aes(x=Publication.Year)) +
  geom_freqpoly(binwidth=10)+
  scale_x_continuous(limits = c(1900, 2030))
```

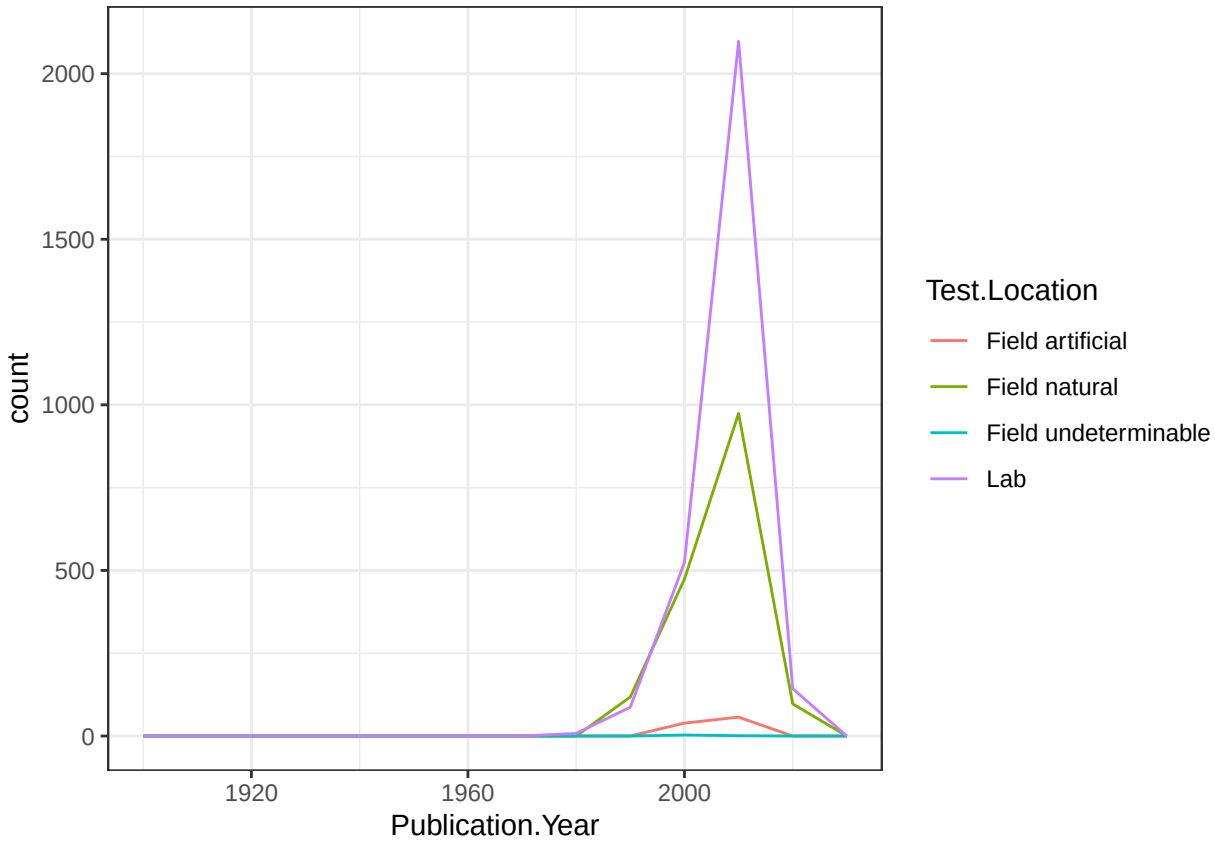
```
## Warning: Removed 2 rows containing missing values (`geom_path()`).
```



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

```
ggplot(Neonics, aes(x=Publication.Year , color=Test.Location)) +
  geom_freqpoly(binwidth=10)+
  scale_x_continuous(limits = c(1900, 2030))+
    theme_bw()
```

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



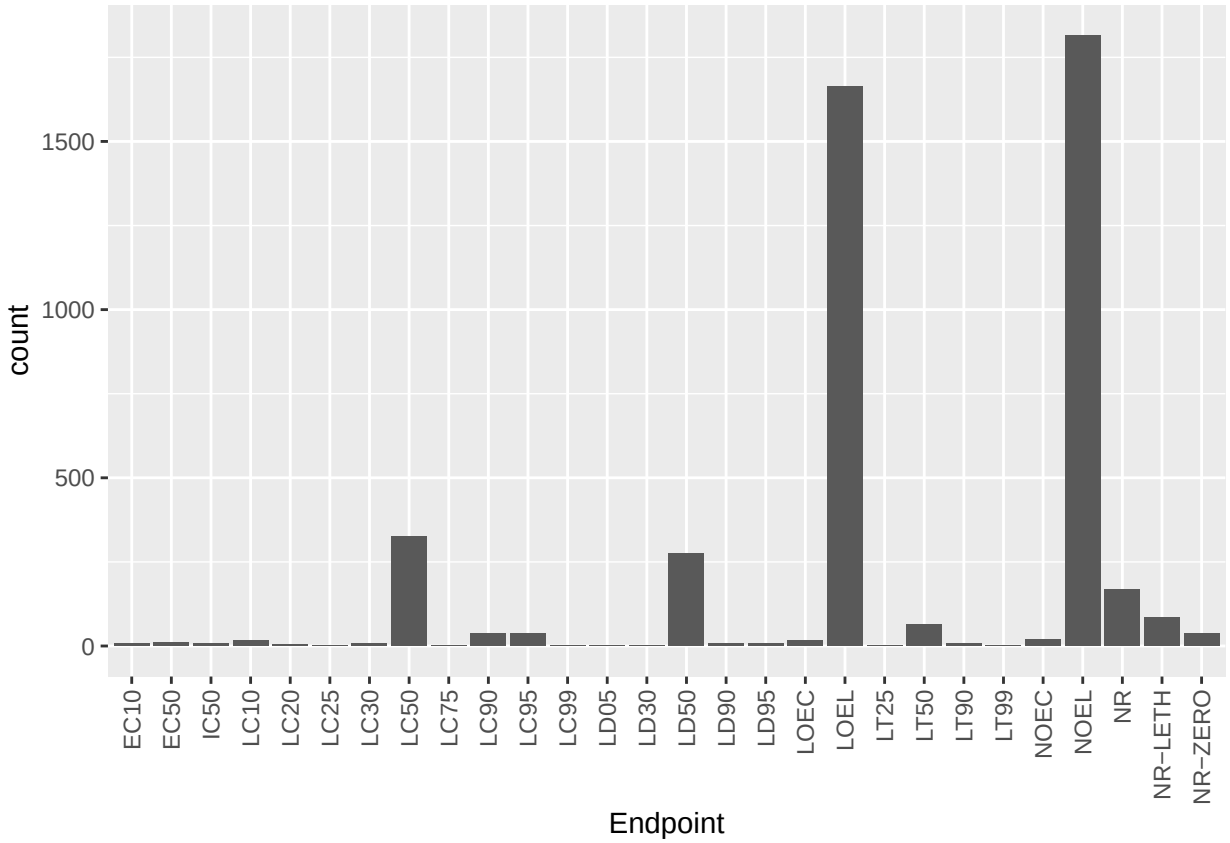
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are in the lab. Testing increases over time across the different locations, though seems to peak in the early 2000s and then experiences a decline.

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 endpoints are NOEL + LOEL. They are defined as Terrestrial.

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)
```

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

```
#it is a factor, not a date

Updated.collectDate <- ymd(Litter$collectDate)
#Updated.collectDate is now updated to the date class

class(Updated.collectDate)
```

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

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
```

```
#[1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051 NIWO_058 NIWO_046
#[11] NIWO_062 NIWO_057
#12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061 ... NIWO_067
```

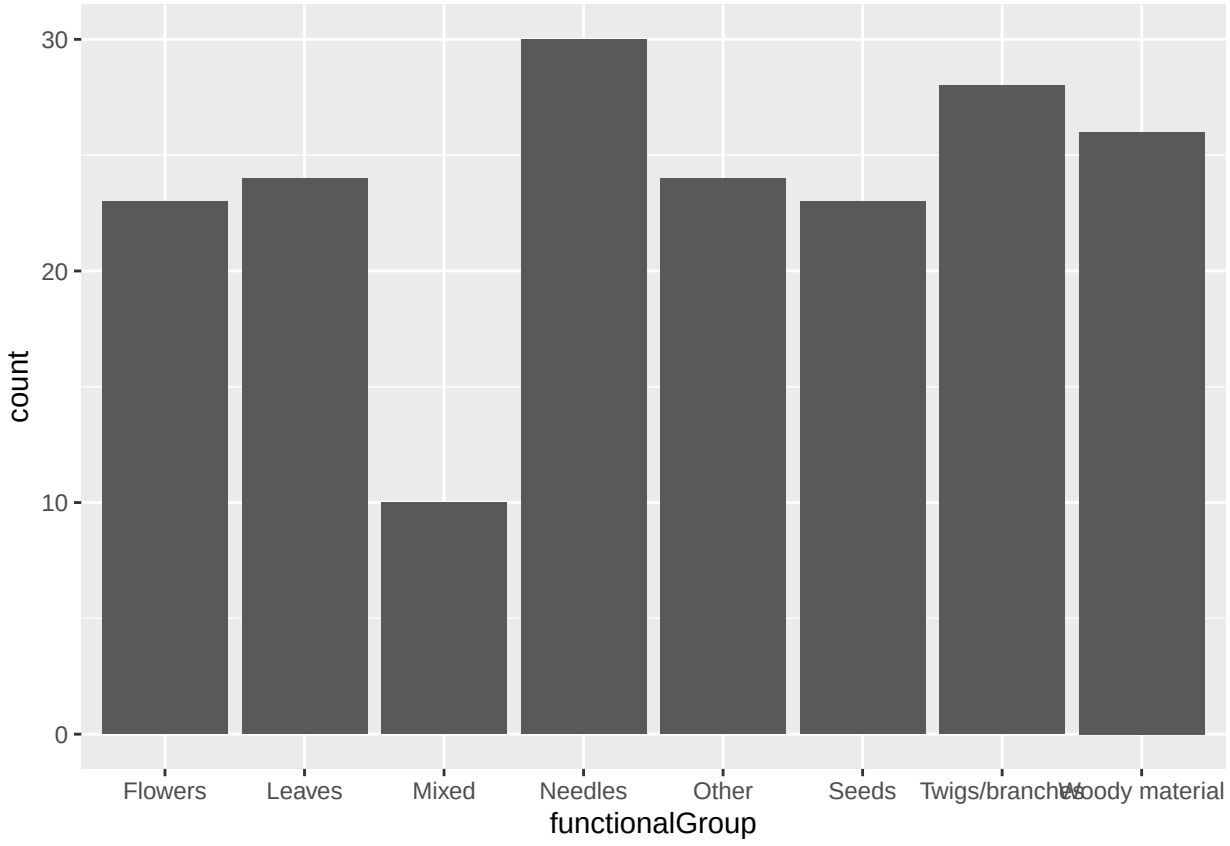
```
summary(Litter$plotID)
```

```
## 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: There are 12 plots. The 'unique' function does not include duplicates, where as the data obtained from 'summary' doesn't get rid of duplicates so all information is there regardless if a repeate.

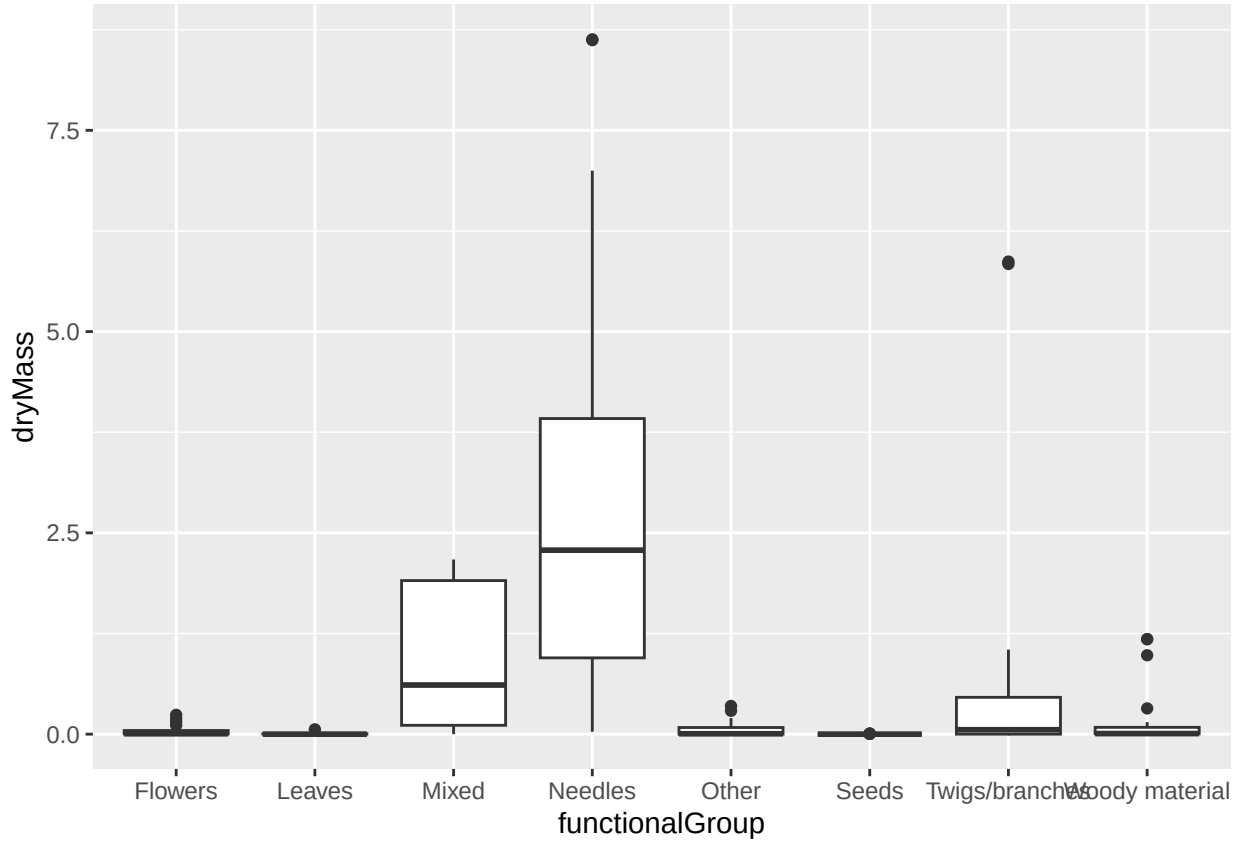
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

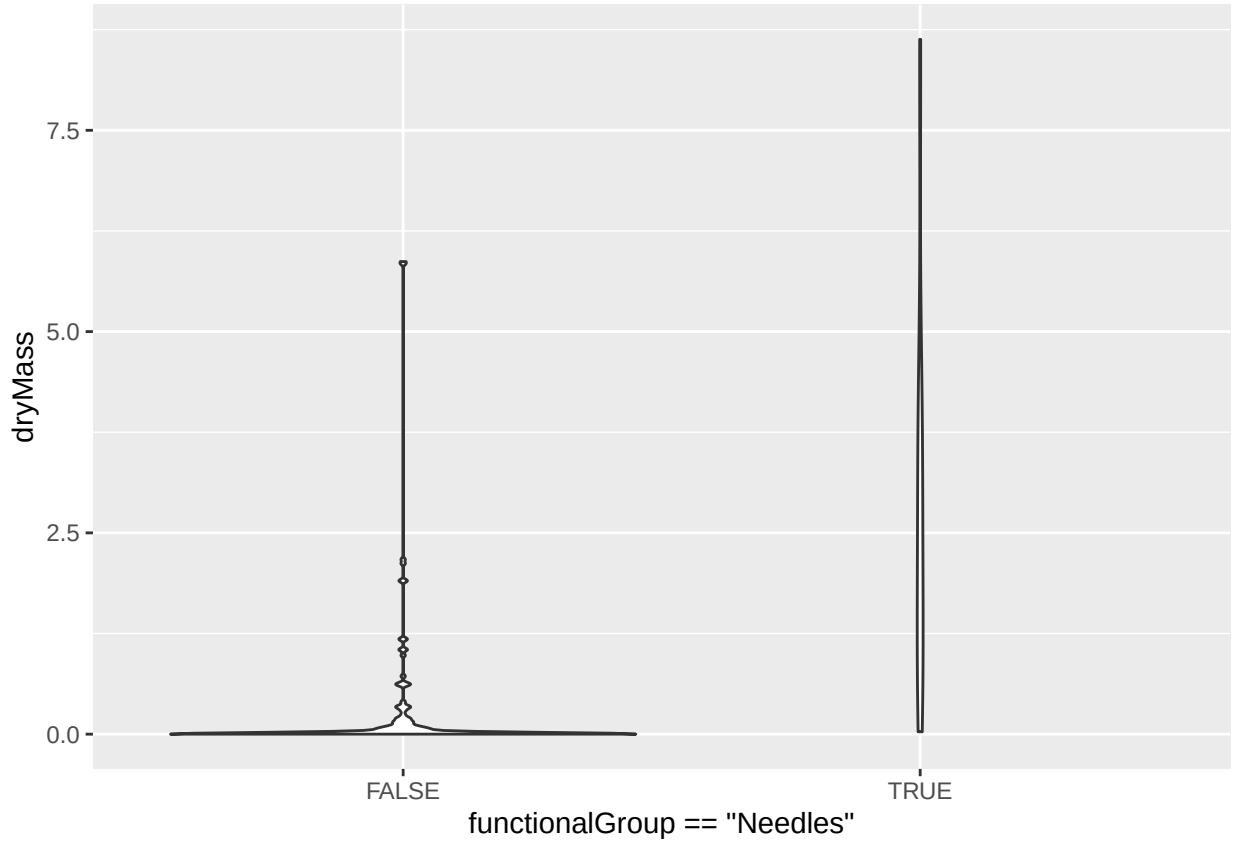


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




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



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

> Answer: The violin plot is very hard to interpret without zooming/adjusting the view, and it is unclear what mass is being represented. The boxplot is well labeled and more obviously interprested given the distribution.

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

> Answer: Needles tend to have the highest biomass at these sites.