Assignment 3: Data Exploration

Cammie Moore

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.

```
# loading libraries
library(tidyverse)
library(lubridate)
library(here)
```

```
## Warning: package 'here' was built under R version 4.4.2
```

```
# checking working directory
getwd()
```

```
## [1] "C:/Users/cammi/OneDrive/Documents/EDE_Spring2025"
```

```
here()
```

```
## [1] "C:/Users/cammi/OneDrive/Documents/EDE_Spring2025"
```

```
# reading in data
insects.df <- read.csv(here("Data", "Raw", "ECOTOX_Neonicotinoids_Insects_raw.csv"),
                       stringsAsFactors = TRUE)
litter.df <- read.csv(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: Neonicotinoids use nicotine to harm insects as a pesticide. Analyzing the effects of the pesticide on organisms is important to analyzing their overall effects on the ecosystems we are responsible for stewarding and human health.

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: Woody debris and litter are vital habitats for various insects and spiders. Researching their locations and impact on insects is important for ecosystem health.

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. Temporal Sampling (once a year) of raised ground traps. 2. Spatial data is collected by mapping each temporal sampling. 3. The annual production is found by collecting the amount of litter from the traps each year.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
str(insects.df)
```

```
## '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'-cy
##  $ 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 ..
##  $ Organism.Lifestage              : Factor w/ 20 levels "Adult","Cocoon",..: 1 1 19 19 19 1 19 1 1
##  $ Organism.Age                    : Factor w/ 39 levels "~10","~24","~7",..: 39 39 39 39 39 36 39 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
##  $ Media.Type                      : Factor w/ 10 levels "Agar","Artificial soil",..: 7 7 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 "~10","~30/","~40/",..: 639 510 813 622 44
##  $ 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
##  $ 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
##  $ Observed.Duration..Days.        : Factor w/ 361 levels "~.1458","~10",..: 145 145 145 145 145 145
##  $ Observed.Duration.Units..Days.  : Factor w/ 17 levels "Day(s)","Day(s) post-emergence",..: 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 NR - NR | Organism Age: \xca
```

```
# 4623 observations of 30 variables

str(litter.df)
```

```
## 'data.frame':    188 obs. of  19 variables:
##  $ uid                     : Factor w/ 188 levels "028eea3d-5c20-4afc-bb7e-a05bab305152",..: 84 96 85
##  $ namedLocation           : Factor w/ 12 levels "NIWO_040.basePlot.ltr",..: 8 8 8 8 8 8 8 8 11 11 ..
##  $ domainID                : Factor w/ 1 level "D13": 1 1 1 1 1 1 1 1 1 1 ...
##  $ siteID                  : Factor w/ 1 level "NIWO": 1 1 1 1 1 1 1 1 1 1 ...
##  $ plotID                  : Factor w/ 12 levels "NIWO_040","NIWO_041",..: 8 8 8 8 8 8 8 8 11 11 ...
##  $ trapID                  : Factor w/ 12 levels "NIWO_040_205",..: 8 8 8 8 8 8 8 8 11 11 ...
##  $ weighDate               : Factor w/ 2 levels "2018-08-06","2018-09-05": 1 1 1 1 1 1 1 1 1 1 ...
##  $ setDate                 : Factor w/ 2 levels "2018-07-05","2018-08-02": 1 1 1 1 1 1 1 1 1 1 ...
##  $ collectDate             : Factor w/ 2 levels "2018-08-02","2018-08-30": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ovenStartDate           : Factor w/ 2 levels "2018-08-02T21:00Z",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ ovenEndDate             : Factor w/ 2 levels "2018-08-06T18:02Z",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ fieldSampleID           : Factor w/ 23 levels "NEON.LTR.NIWO040205.20180802",..: 14 14 14 14 14 14
##  $ massSampleID            : Factor w/ 168 levels "NEON.LTR.NIWO040205.20180802.FLR",..: 102 101 103 9
##  $ samplingProtocolVersion: Factor w/ 1 level "NEON.DOC.001710vE": 1 1 1 1 1 1 1 1 1 1 ...
##  $ functionalGroup         : Factor w/ 8 levels "Flowers","Leaves",..: 7 6 8 1 8 4 5 2 1 8 ...
##  $ dryMass                 : num  0.4 0.005 0.04 0.005 0.07 1 0.2 0.005 0.19 1.18 ...
##  $ qaDryMass               : Factor w/ 2 levels "N","Y": 1 1 2 1 1 1 1 1 1 2 ...
```

```
##  $ remarks                : logi  NA NA NA NA NA NA ...
##  $ measuredBy             : Factor w/ 2 levels "kstyers@battelleecology.org",..: 1 1 1 1 1 1 1 1 1 1
```

```
# 188 observations of 19 variables
```

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...]

```
sort(summary(insects.df$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: The population (1803) and mortaility (1493) are the two highest effects. These would be of interest because insects in great numbers are experiencing varying population effects from the Neonicotinoids or mortality.

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...]

```
summary(insects.df$Species.Common.Name, maxsum = 6)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                       285                  183
##   Carniolan Honey Bee            Bumble Bee               (Other)
##                   152                       140                 3196
```

Answer: Honey Bee, Parasitic Wasp Buff, Tailed Bumbled Bee, Carniolan Honey Bee, Bumble Bee, Other

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...]

```
# view(insects)
```

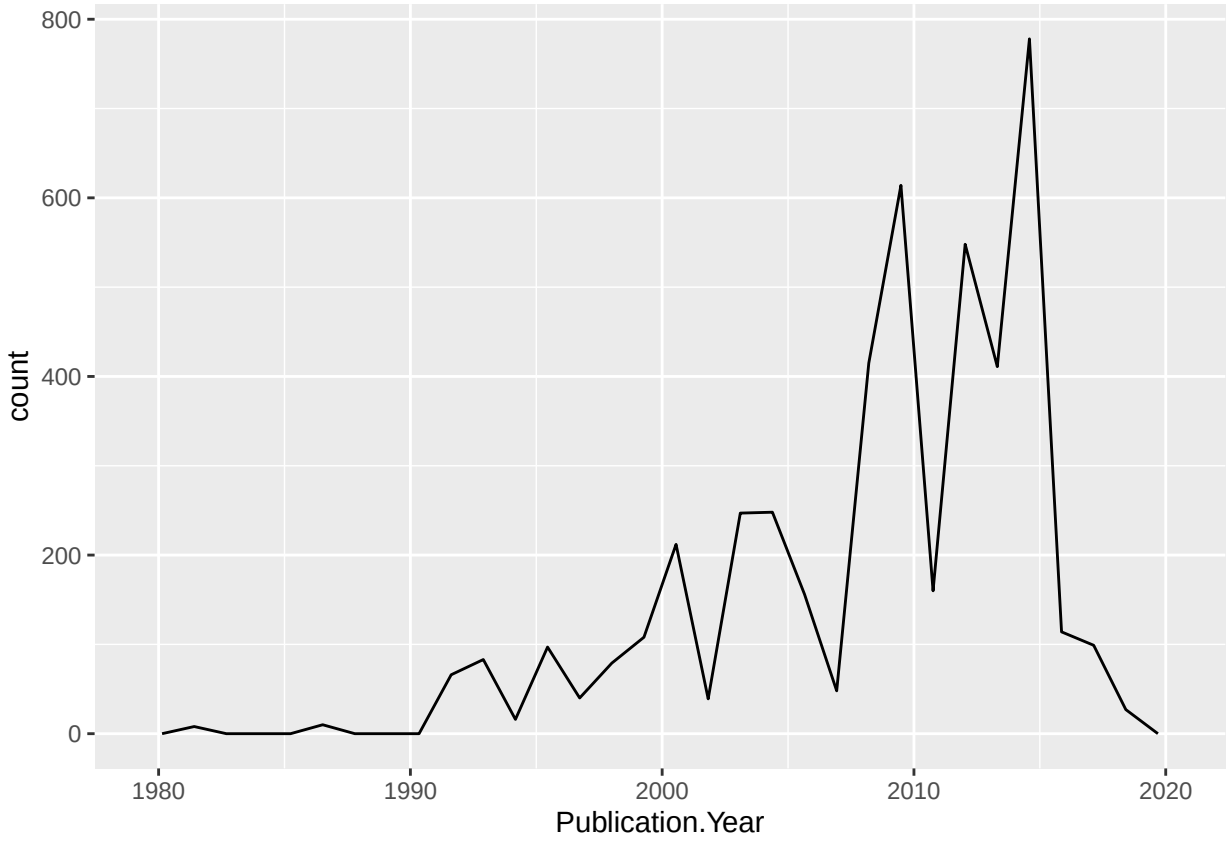
Answer: The Conc.1.Author are factor variables. I believe it is showing the amount of the active ingredient. It is better to have thisdata as categorical rather numeric because the groups are set up as less than x, therefore making the numbers more categorical than numeric.

Explore your data graphically (Neonics)

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

```
ggplot(data = insects.df)+
  geom_freqpoly(aes(Publication.Year))
```

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

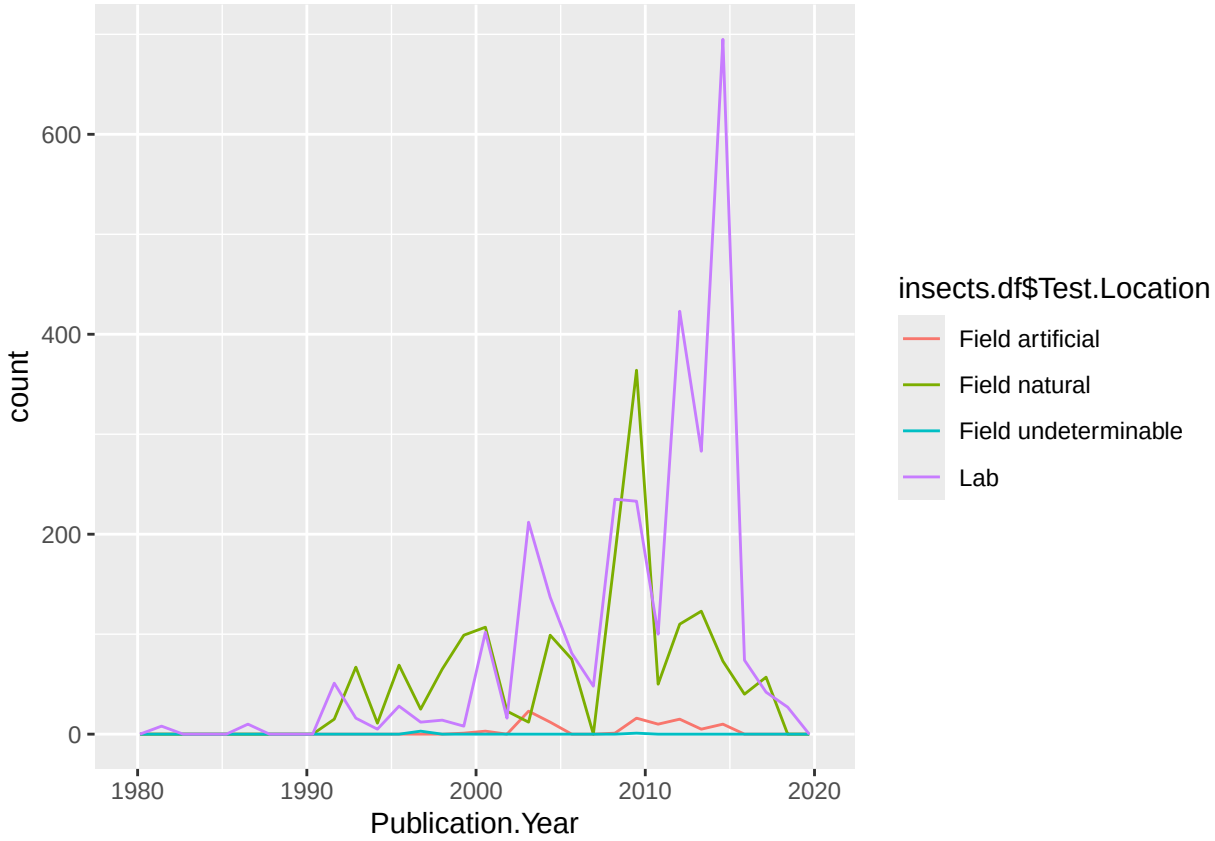


10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
ggplot(data = insects.df)+
  geom_freqpoly(aes(Publication.Year, color = insects.df$Test.Location, group = insects.df$Test.Locatio
```

```
## Warning: Use of `insects.df$Test.Location` is discouraged.
## i Use `Test.Location` instead.
## Use of `insects.df$Test.Location` is discouraged.
## i Use `Test.Location` instead.
```

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



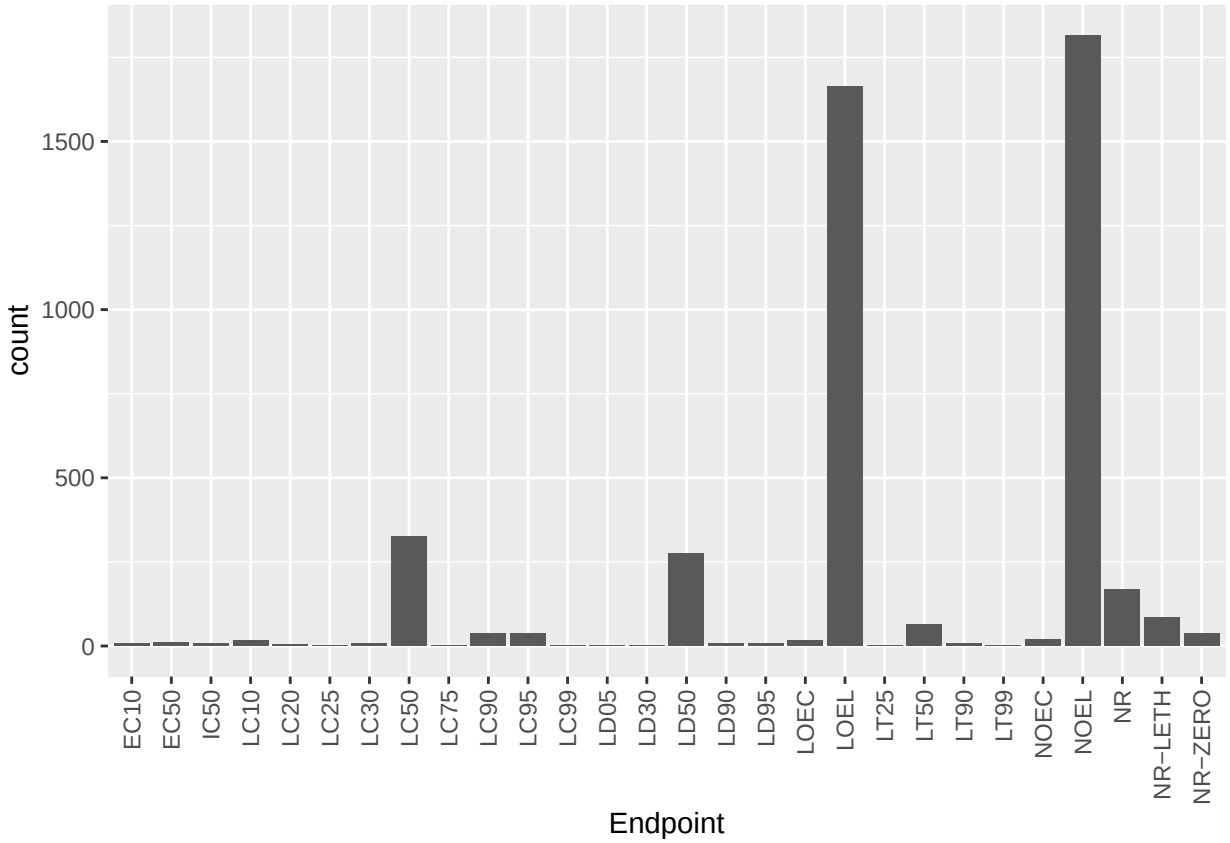
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations appear to be natural fields and labs. Both increased over time, but beginning in 2010 lab testing rapidly overtook natural field testing.

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



Answer: The most common two endpoints are LOEL and NOEL. NOEL is defined as the chemical having No-observable-effect-level. The highest dose (concentration) producing effects did not significantly differ from the responses of the controls. LOEL is the lowest-observable-effect-level. It is the lowest dose (concentration) which produces effects that are significantly different from the responses of the 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.

```
unique(litter.df$collectDate) # not a date, a factor
```

```
## [1] 2018-08-02 2018-08-30
## Levels: 2018-08-02 2018-08-30
```

```
litter.df$collectDate <- ymd(litter.df$collectDate)
# litter was collected on August 2nd, 2018 and August 30, 2018
```

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`?

```
unique(litter.df$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
```

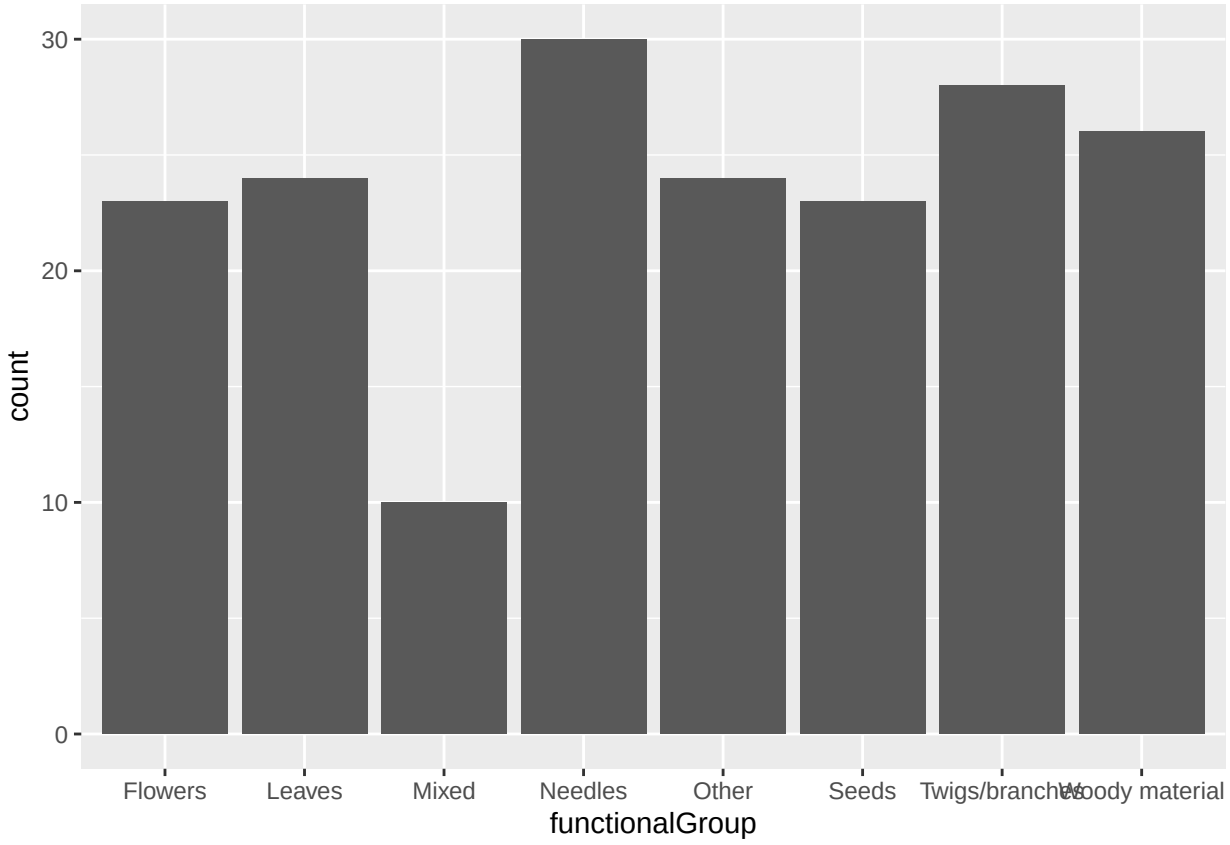
```
summary(litter.df$plotID)
```

```
## 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. Unique only identifies the number of unique values in a vector, while summary identifies how many data points fall into each unique value in a vector.

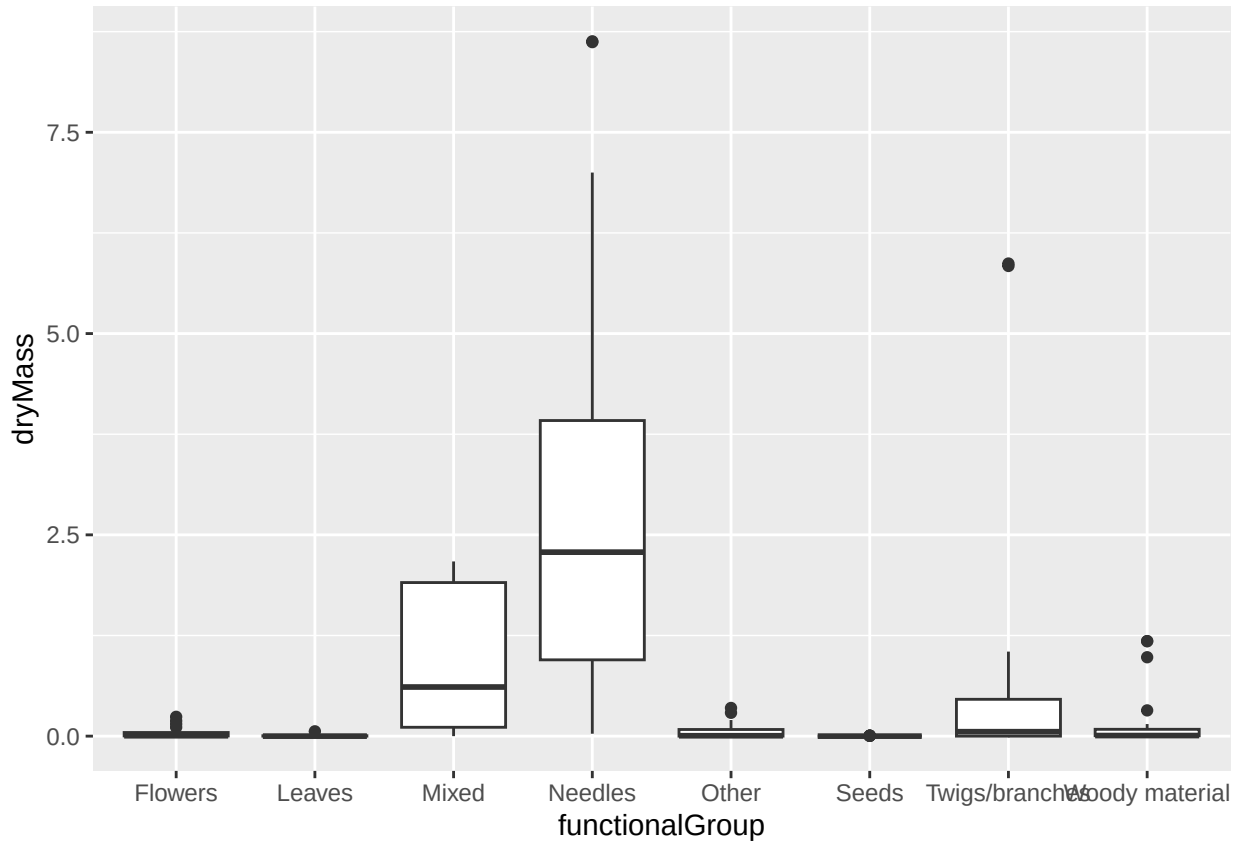
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(data = litter.df)+
  geom_bar(aes(functionalGroup))
```

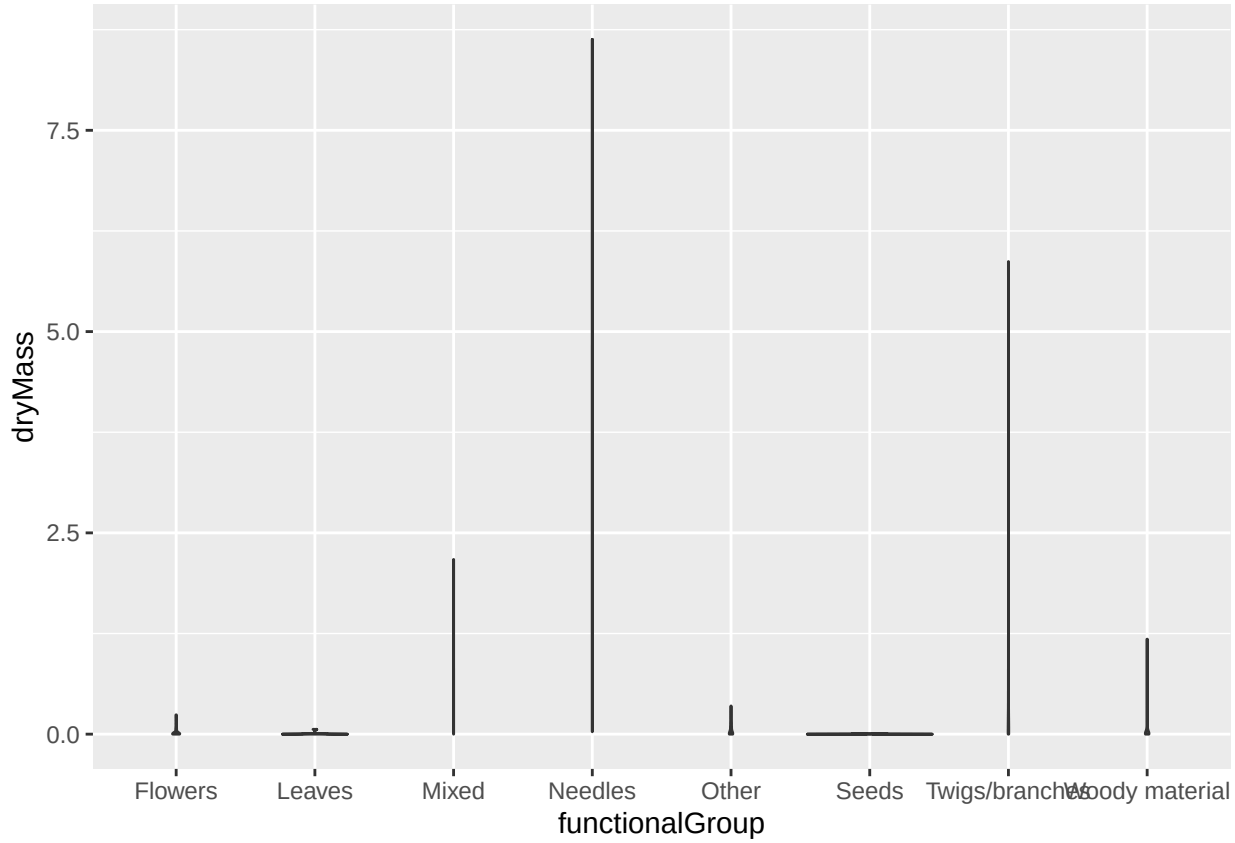


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(data = litter.df)+
  geom_boxplot(aes(functionalGroup, dryMass))
```




```
ggplot(data = litter.df)+
  geom_violin(aes(functionalGroup, dryMass))
```



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

> Answer: The boxplot is a much more effective visualization than the violin plot for this case. The violin plot because it seems there is not a normal distribution of weight for most materials.

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

> Answer: Needles, mixed, and twigs and branches appear to have the greatest amount of biomass at these sites.