Assignment 3: Data Exploration

Emily Wood

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.

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] "/home/guest/EDA_2022/EDA-Fall2022"
```

```
#install.packages("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)

knitr::opts_chunk$set(tidy.opts=list(width.cutoff=80), tidy=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: We would be interested in the ecotoxicology of neonicotinoids on insects to see the results of the pesticides. Were the pesticides effective, were there any unforeseen consequences such as behavioral changes or emergence patterns. It is also important to know how these chemicals affect insects that were not the target pest.

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 would give insight to how the organic material of soil is formed over time. It also shows the composition for habitat for numerous species including insects, invertebrates and other smaller organisms. Litter and debris also act as a barrier between soil and the elements. It prevents the soil itself from drying out or eroding away during precipitation events.

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 traps and ground traps. 2.All masses measured are reported at the spatial resolution of a single trap and the temporal resolution of a single collection event. 3. Mass data for each collection event are measured separately for different groups. Examples of these groups include leaves, twigs, Needles, Seeds, etc.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
length(Neonics)
```

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

```
dim(Neonics)
```

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

```
colnames(Neonics)
```

```
##  [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"
```

```
str(Neonics)
```

```
## '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 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 1
##  $ Organism.Age                    : Factor w/ 39 levels "<=24","<=48",..: 39 39 39 39 39 36 39 36 36
##  $ 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
```

```
# There are 30 columns and 4,623 entries in the Neonics dataframe.
```

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 most commonly studied effects are mortality and population. It makes sense that mortality is studied as that is often the desired outcome of pesticides. Population also makes sense especially if a pest is overpopulated in an area. By studying population you can assess how it fluctuates based on the use of this pesticide in an area.

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: The most commonly studied species is the Honey Bee and two types of bumble bees. They are likely studied because they are important crop and native plant pollinators. If these pesticides are killing them at a high enough rate this could cause a collapse in our food system.

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"
```

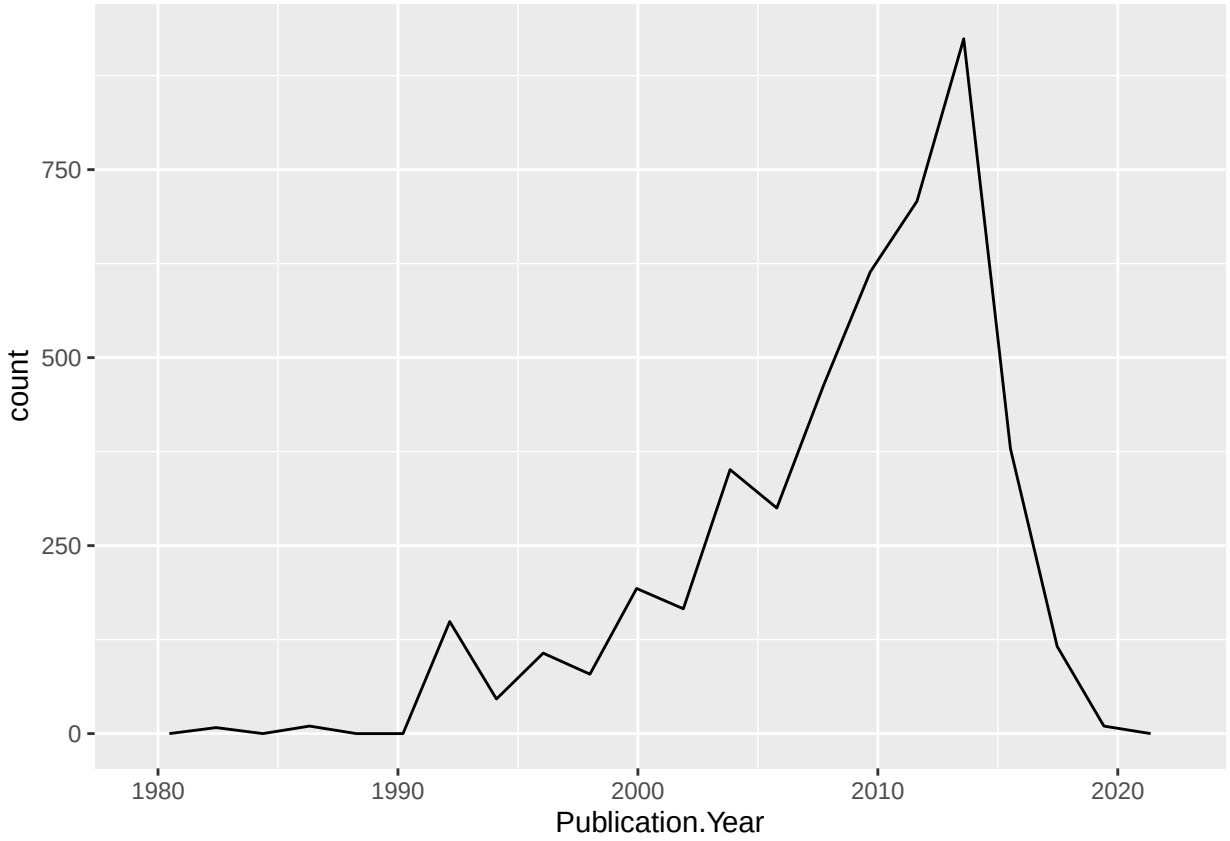
Answer: It is not currently numeric because we set everything as a factor when we read in the csv file.

Explore your data graphically (Neonics)

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

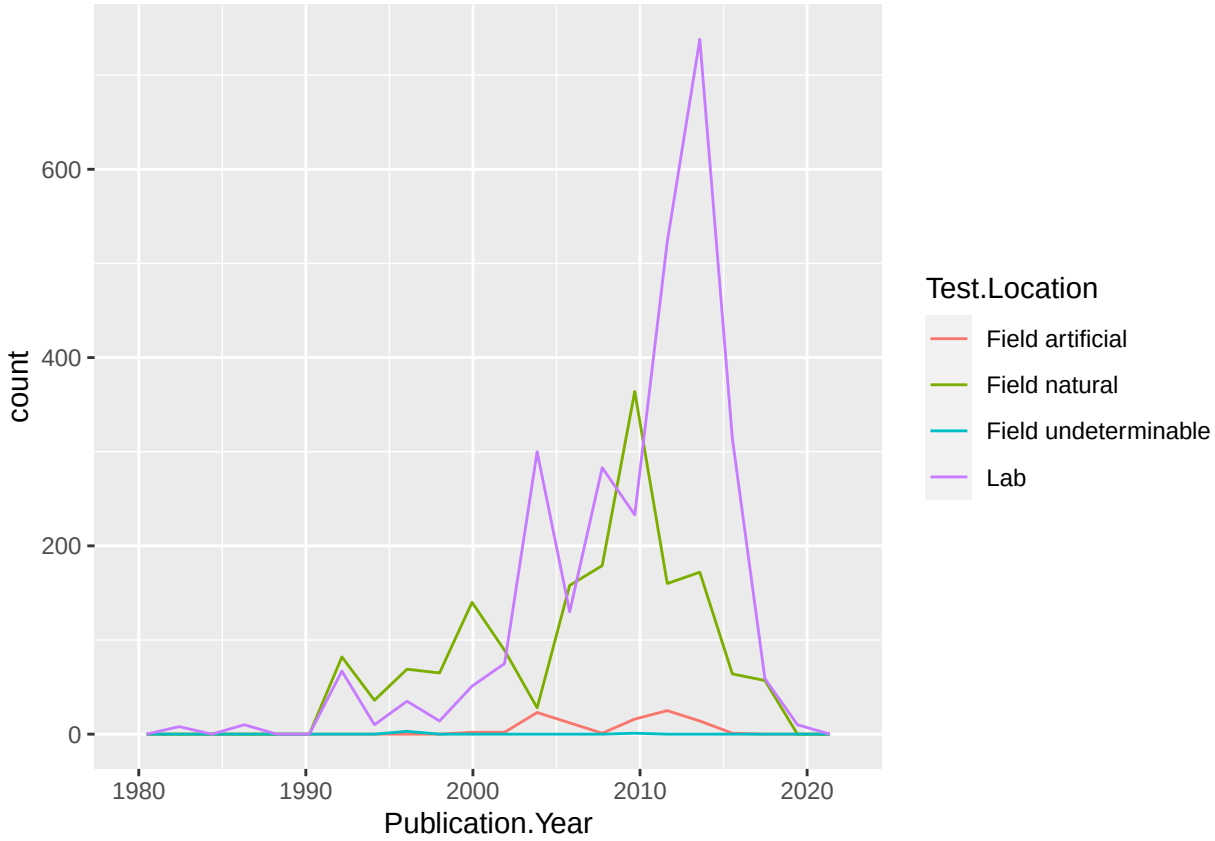
```
library("ggplot2")

ggplot(Neonics) + geom_freqpoly(aes(Publication.Year), bins = 20)
```



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),
    bins = 20)
```

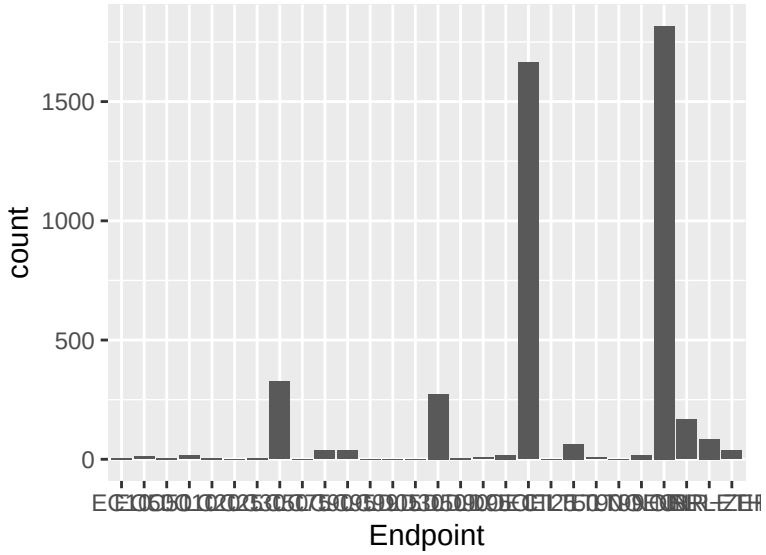


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are "Field Natural" and "Lab". According to the graph the "Field Natural" test location was slightly more common through the 90s. In the early 2000s the Lab became the most common test location with "field Natural" spiking one more time in 2010.

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)) + geom_bar()
```




```
summary(Neonics$Endpoint)
```

```
##     EC10     EC50     IC50     LC10     LC20     LC25     LC30     LC50     LC75     LC90
##        6       11        6       15        5        1        6      327        1       37
##     LC95     LC99     LD05     LD30     LD50     LD90     LD95     LOEC     LOEL     LT25
##       36        2        1        1      274        6        7       17     1664        1
##     LT50     LT90     LT99     NOEC     NOEL       NR  NR-LETH  NR-ZERO
##       65        7        2       19     1816      167       86       37
```

Answer: The two most popular endpoints are NOEL and LOEL. NOEL stands for no observable effect level. The highest concentration does not show effects that differ from control groups. LOEL stands for lowest observable effect level. This means that the lowest dose producing effects were significantly different from 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.

```
class(Litter$collectDate)
```

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

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

unique(Litter$collectDate, 2018 - 8)
```

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

```
# Litter was samples twice in AUGUST OF 2018. The collection dates were on the
# 2nd and 30th.
```

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)
```

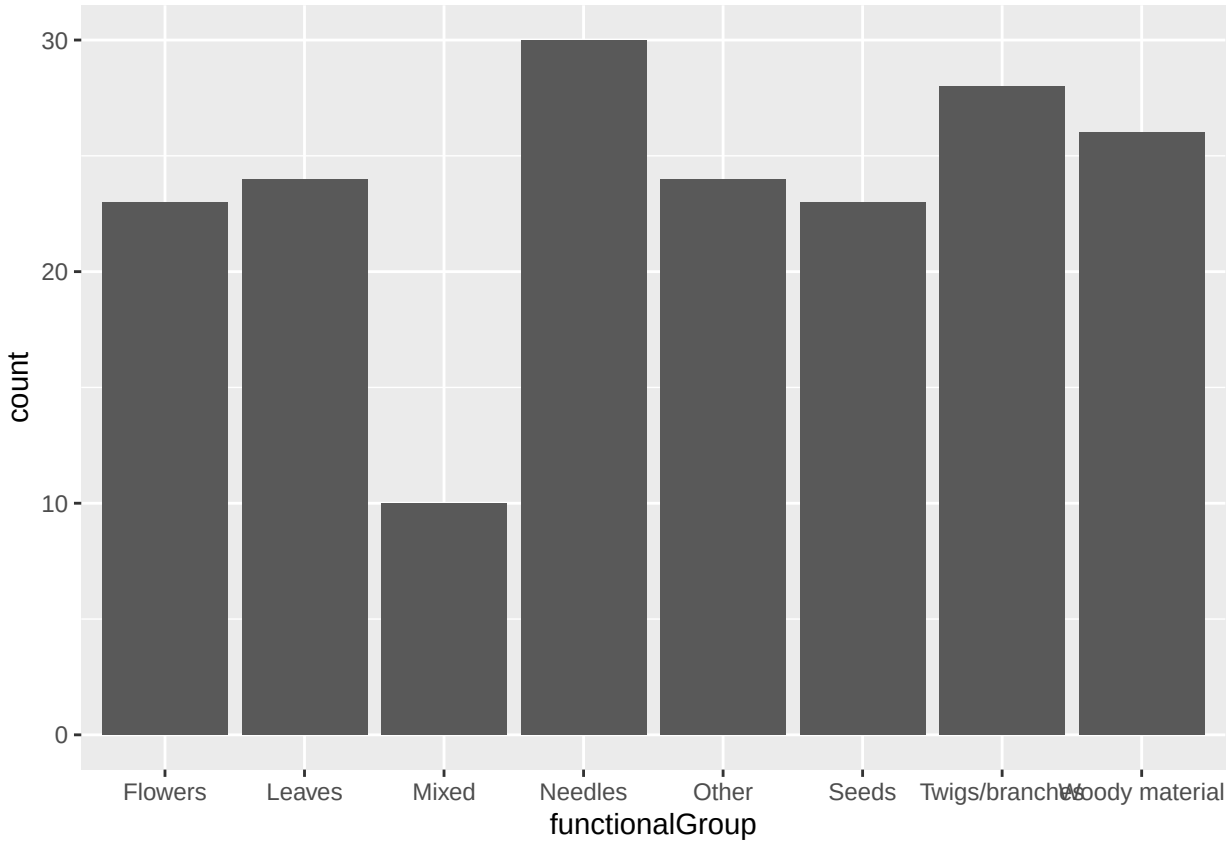
```
##  [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
```

```
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

Answer: 12 plots were sampled at Niwot Ridge. The unique function gives us a count of how many variables are in the vector. The summary function tells us how many times each variable is listed in the vector.

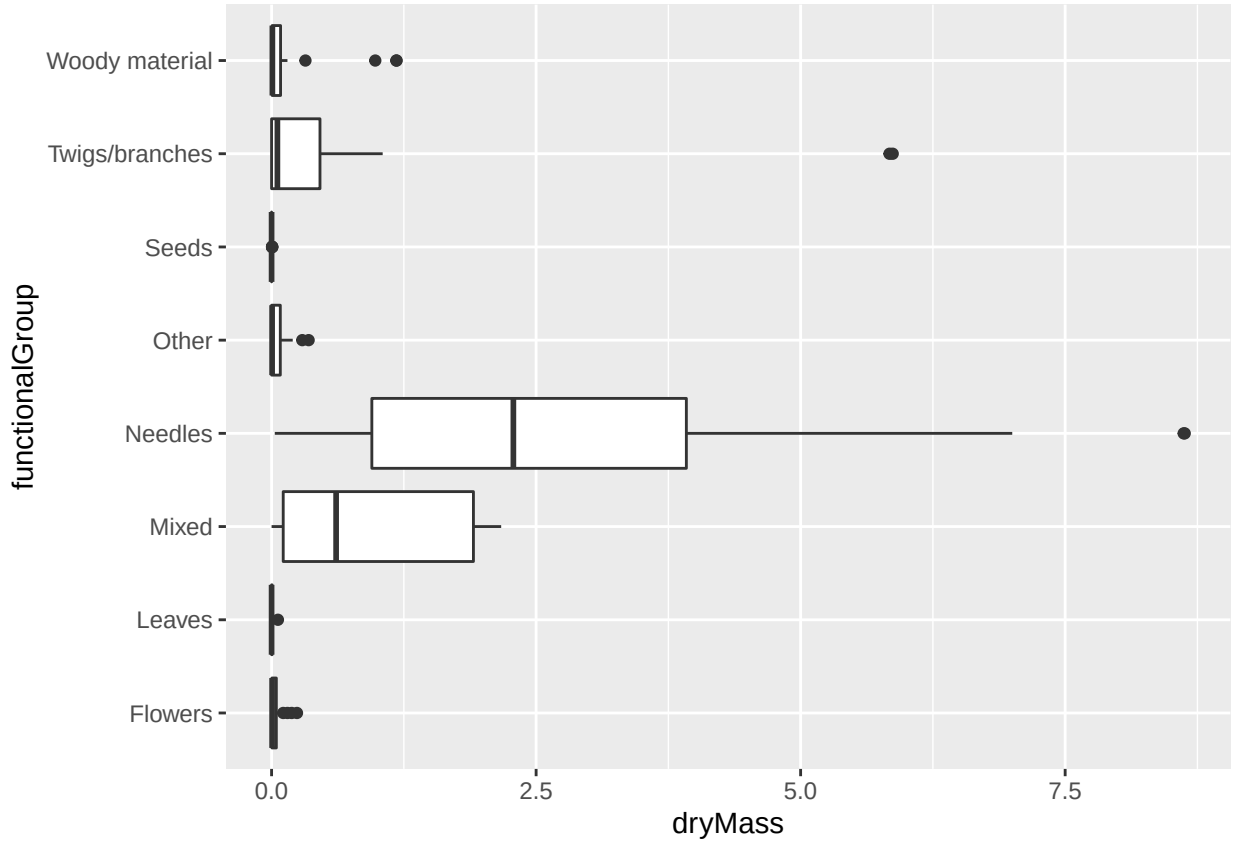
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

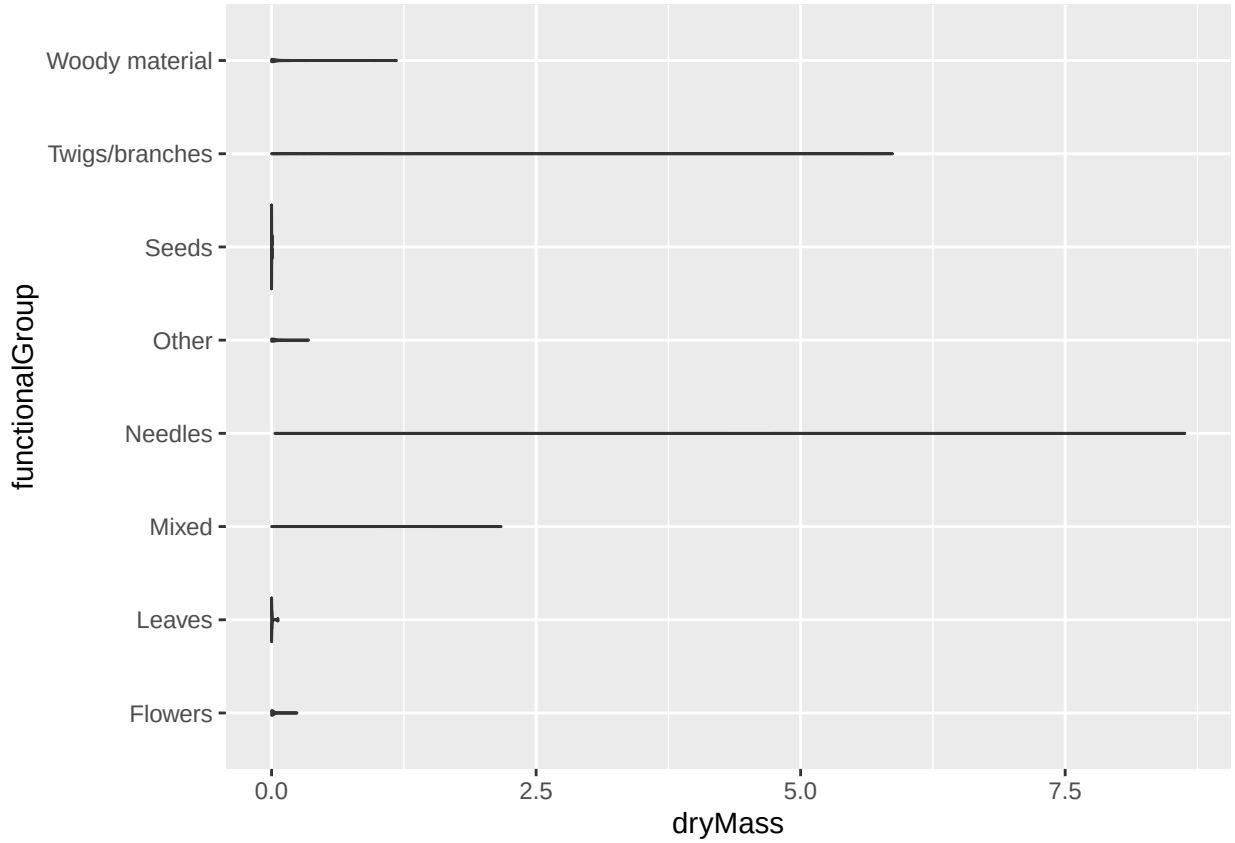


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




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



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

> Answer: Boxplots show summary statistics. Violin plots take that one step further and also show density. In this case, the density does not compound enough to show any trends in dryMass. We can see some areas of higher density at 0 for seeds and leaves but there is not enough information to give us the recognizable box plot waves.

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

> Answer:According to the boxplots, needles tend to have the highest biomass at these sites. This is followed by the mixed category and twigs category.