Assignment 3: Data Exploration

Kathleen Mason

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Exploration.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Salk_A03_DataExploration.Rmd") prior to submission.

The completed exercise is due on Tuesday, January 28 at 1:00 pm.

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.

```
getwd()
```

```
## [1] "/Users/kathleenmason/Documents/DUKE/Data Analytics/Environmental_Data_Analytics_2020/Assignments
```

```
library(tidyverse)

Neonics <- read.csv ("../Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv")

Litter<- read.csv ("../Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv")
```

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 ecotoxicologoy of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: We might be interested in it to see if the insecticide (neonicotinoids) are actually having an effect on the insecticide. An insecticide is supposed to kill the insects, so ecotoxicology research may hope to obtain information on if a certain insecticicde works on certain insects or not.

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: Knowing decomposition rates of litter and woody debris in forests may be benefifical information in determining carbon and nutrient storage and general cycling in particular forests.

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: *Longer litter material is captured in ground traps and shorter are captured in elevated traps.* These traps are paired and placed every 400 square meters, leaving about 1-4 pairs in each plot *plots are randomly placed when aerial cover is greater than 50%, and placed heterogeneously when less than 50% aerial cover.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

6. Using the `summary` function, 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: Population effects are studied the most. This might be of particular interest in order to see how a population shifts in response to different factors.

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: They are all different types of bees. Bees are probably of more interest because we don't want to kill off this species, so we need to understand the effect insecticides have on them.

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:factor. It is not numeric because while there are numbers, they must not be values.

Explore your data graphically (Neonics)

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

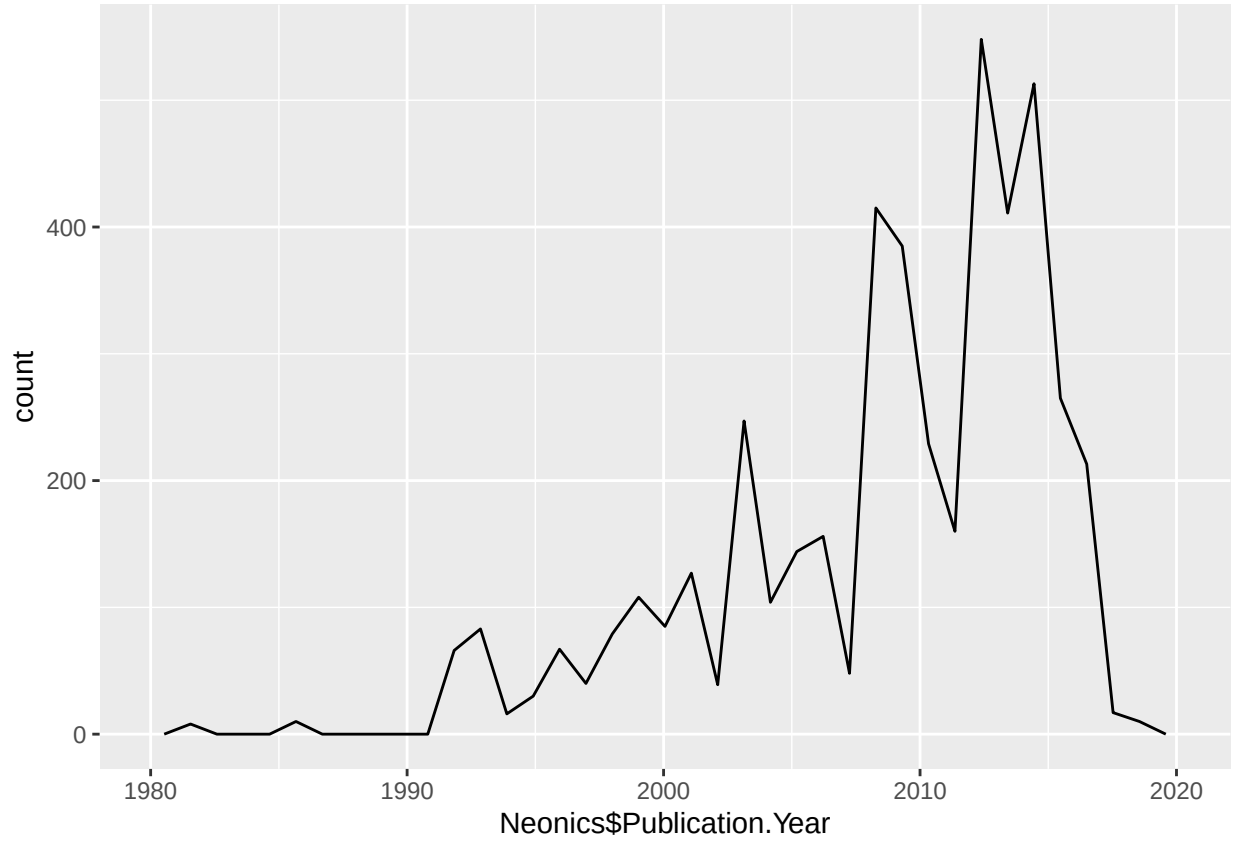
```
class(Neonics$Publication.Year)
```

```
## [1] "integer"
```

```
summary(Neonics$Publication.Year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1982    2005    2010    2008    2013    2019
```

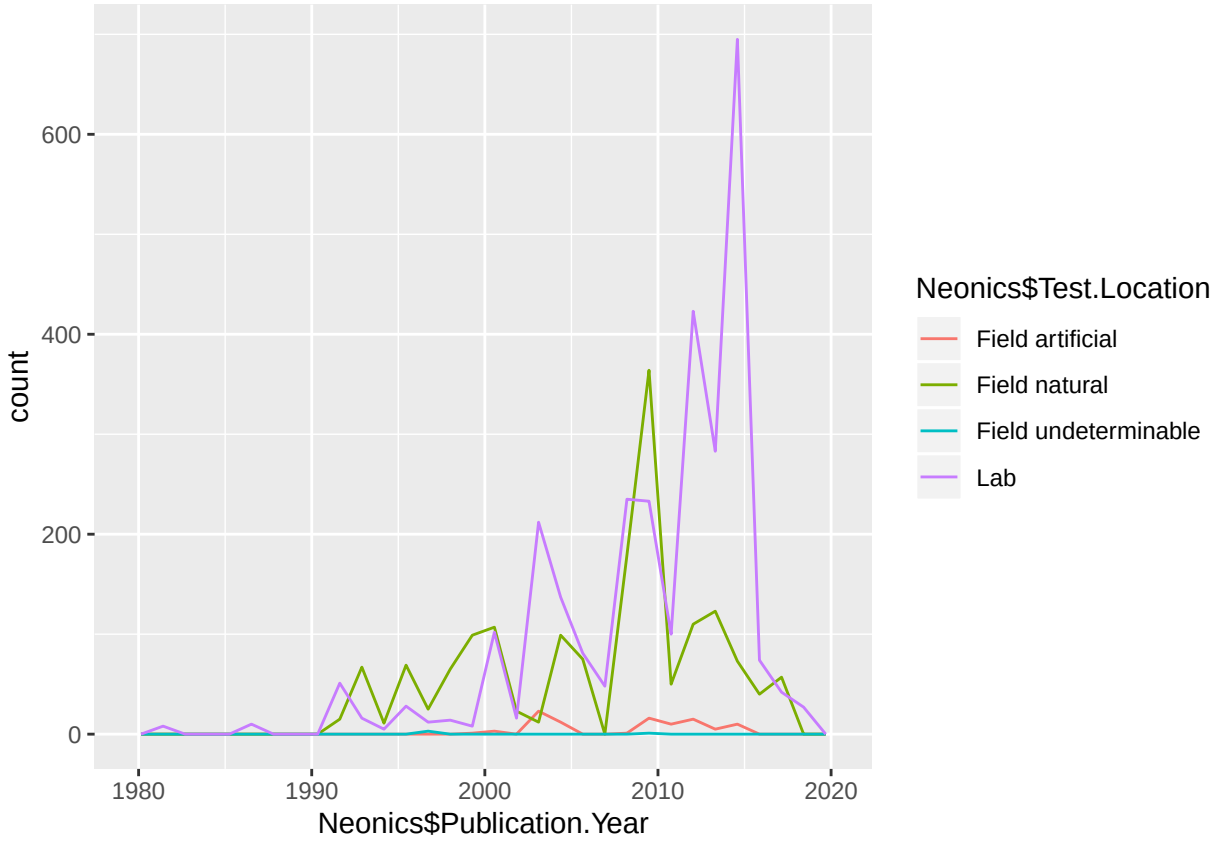
```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Neonics$Publication.Year), bins = 37)
```



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 = Neonics$Publication.Year, color = Neonics$Test.Location))
```

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

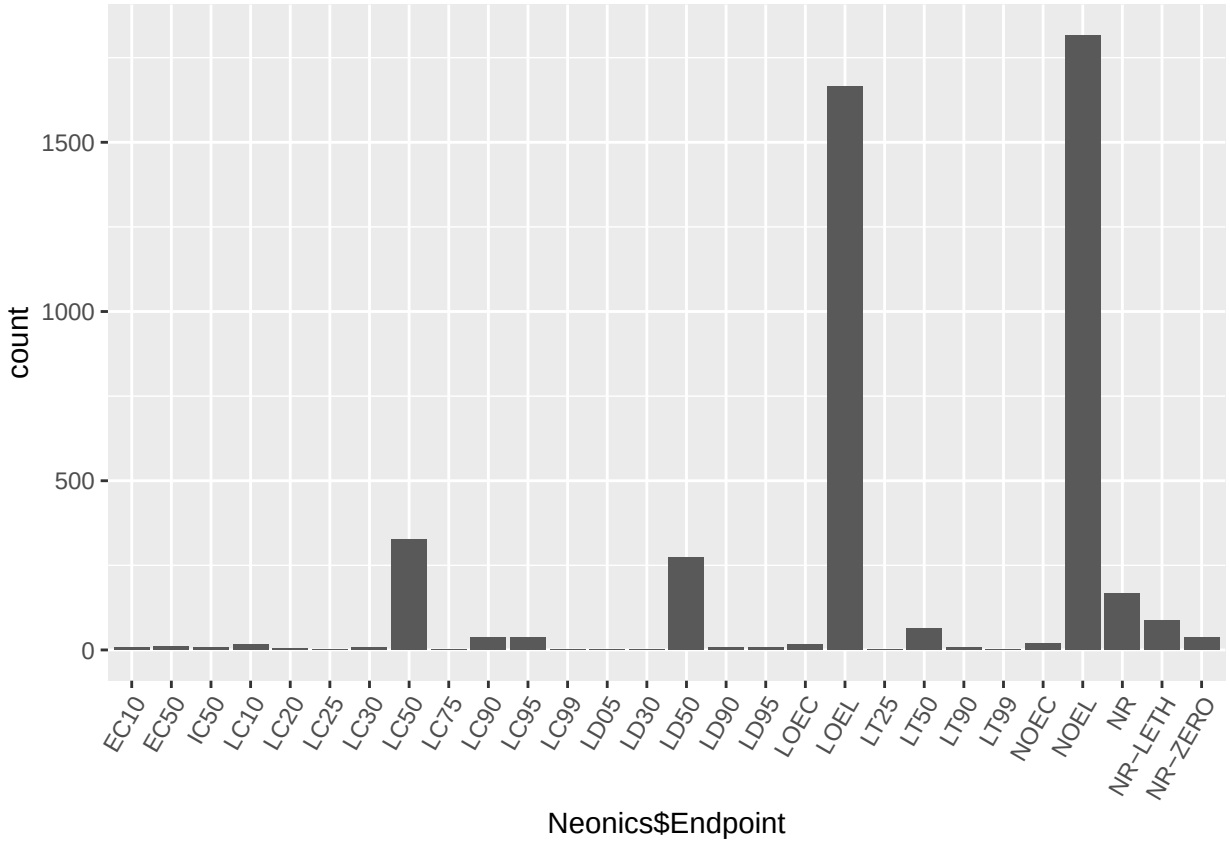


Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: The lab is the most common test location and has been more frequent over time, while most recently, natural field test locations have declined.

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 = Neonics$Endpoint)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```




```
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: LOEL and NOELare the most common end points.LOEL is lowest observable effect and NOEL is no observable effect.

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"
```

```
help(as.Date)
today <- Sys.Date()
format(today, format = "%B")
```

```
## [1] "January"
```

```
format(today, format = "%a")
```

```
## [1] "Sun"
```

```
format(today, format = "%Y")
```

```
## [1] "2020"
```

```
##

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


class(Litter$collectDate)
```

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

```
unique(Litter$collectDate)
```

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

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$siteID)
```

```
## [1] NIWO
## Levels: NIWO
```

```
summary(Litter$siteID)
```

```
## NIWO
##  188
```

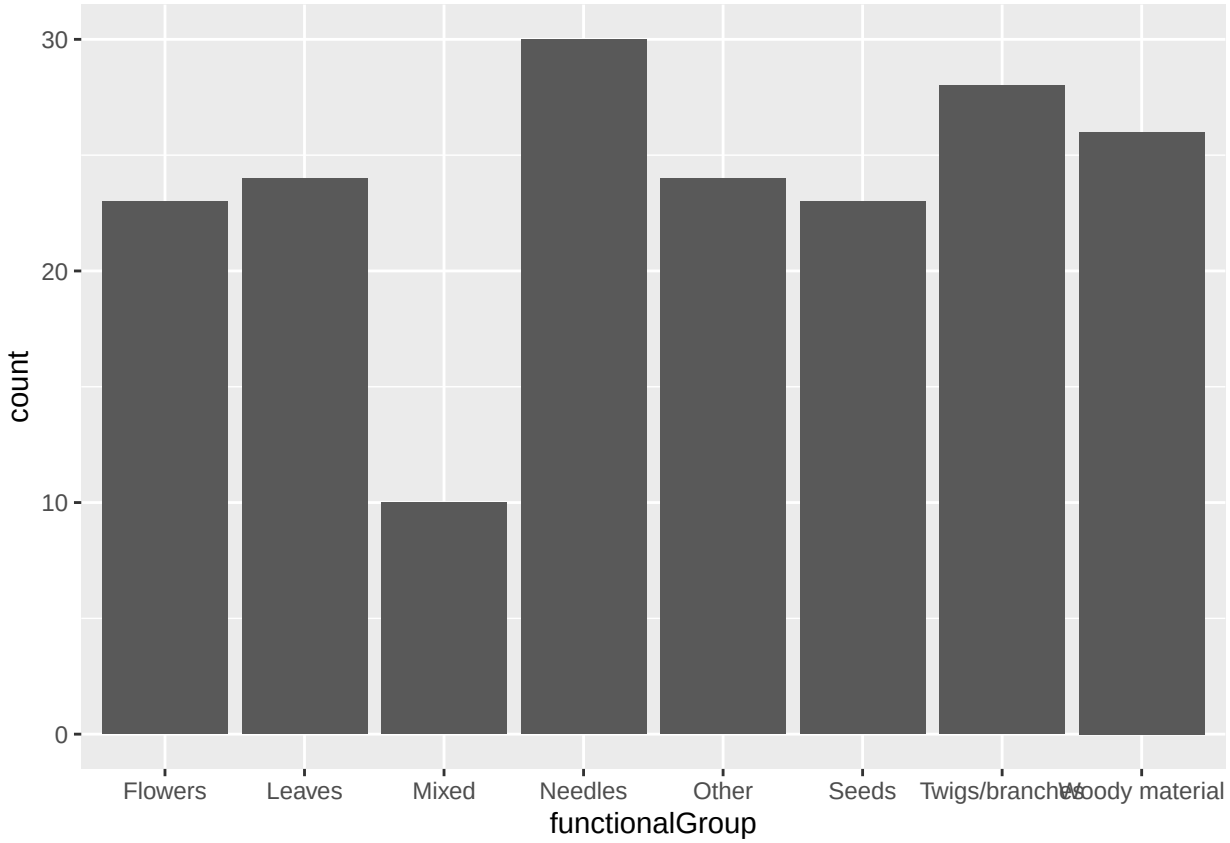
Answer: The unique function outputs the number of levels or different answers in each column, while the summary outputs the different levels and the amount of each.

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.

```
class(Litter$functionalGroup)
```

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

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

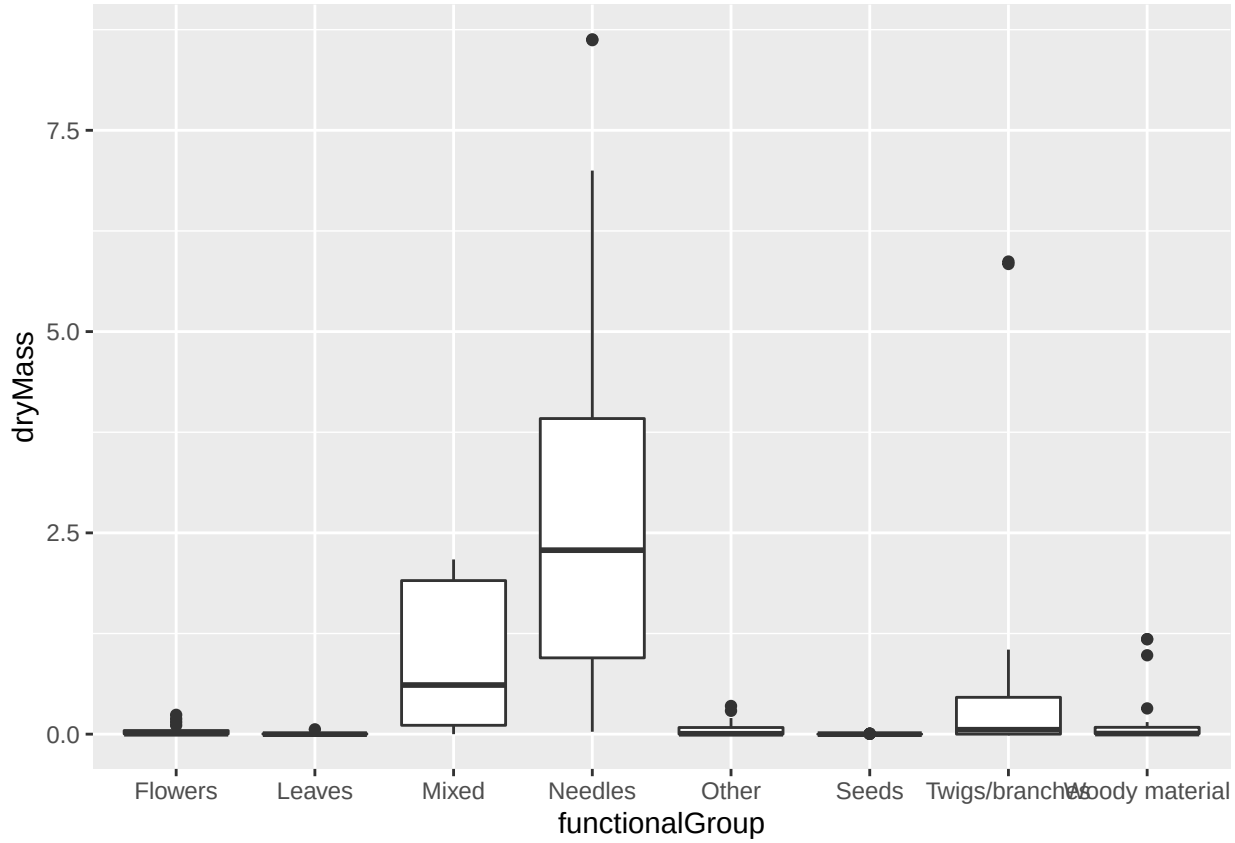


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
class(Litter$dryMass)
```

```
## [1] "numeric"
```

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

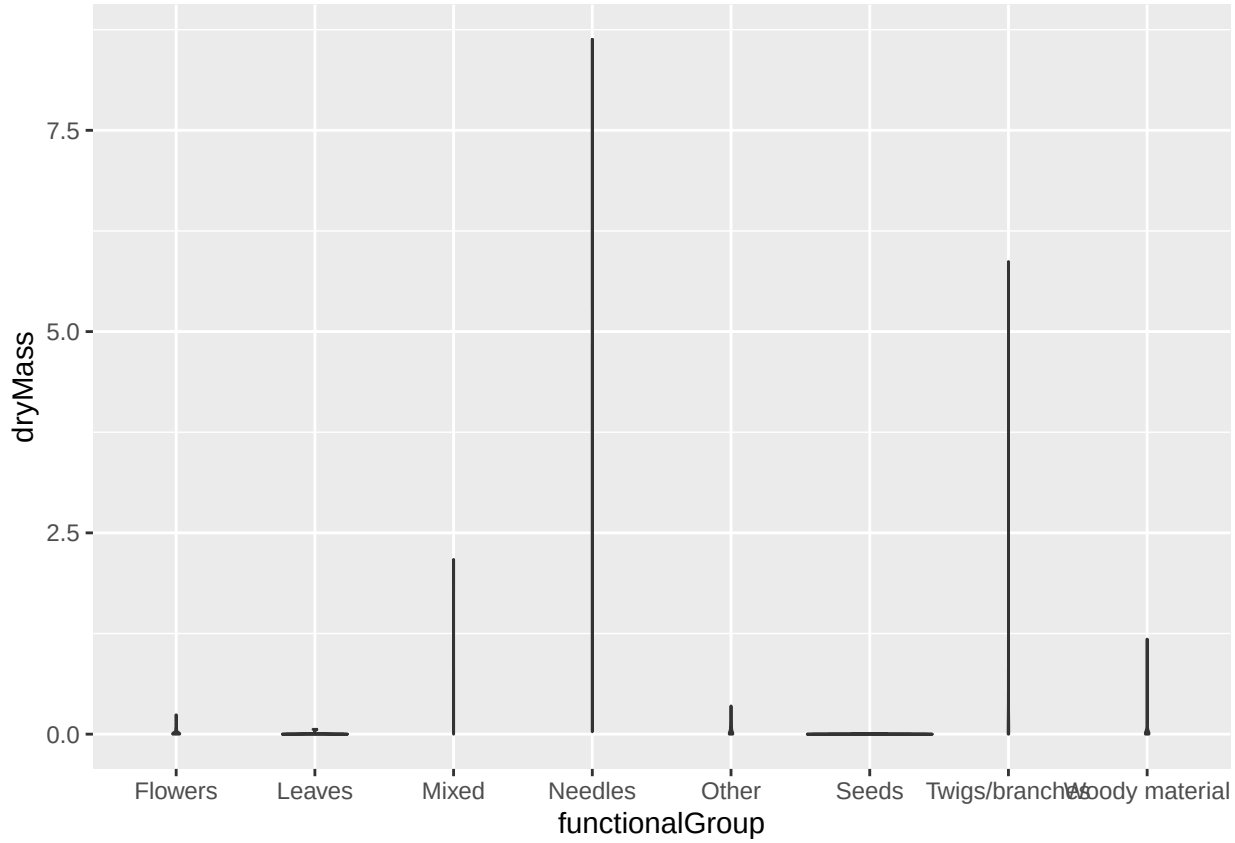



```
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75))
```

```
## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to unique
## 'x' values

## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to unique
## 'x' values

## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to unique
## 'x' values
```



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

> Answer: There aren;t that many samples of each functional group, and the number of samples of functional groups is pretty evenly distributed. Violin might work best if there were much more samples of one group over others. Theboxplot best shows the data we really care about.

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

> Answer: Needles and mixed show the highest biomass.