Assignment 3: Data Exploration

Sena McCrory

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] "C:/Users/senam/Box Sync/My Documents/MEM classes/Duke Spring 2020/DataAnalytics/Environmental_Da
```

```
library(tidyverse)

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

litter.data <- 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: neonics are are widly used category of pesticides used in agricultural production. They have been implicated in the collapse of bee and other important/beneficial insect populations.

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: Litter and woody debris in forests can help us to understand the productivity of the forest and information about carbon and other nutrient cycles including fluxes, decomosition rates, and other biogeochemical measurements. Timing of debris amounts can also tell us about the phenology of leaf senescence and leaf drop in fall.

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:

- masses are reported for separate functional groups - e.g. leaves, twigs, seeds, flowers, etc
- the sampling design is spatially distributed with a pair (one elevated and one ground litter trap) per 400 sq m of the study plot (there are 20 plots), so there may be varying numbers of traps for different plots
- temporal sampling is irregular - with more frequent sampling during autumn and gaps during winter or dormant seasons

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(neonic.data)
```

```
## [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(neonic.data$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 were "mortality" and "population" likely becuase these are the easiest and quickest to test. These effects are of interest because the researchers want to determine whether exposure to neonics could be responsible for decreased insect populations.

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(neonic.data$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 top 6 most reported species are all hymenoptera (bees and wasps) - bees are ecologically important for pollination and therefore food production and parasitic wasps are often beneficial insects for agricultural crops because they control populations of unwanted pest insects

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(neonic.data$Conc.1..Author.)
```

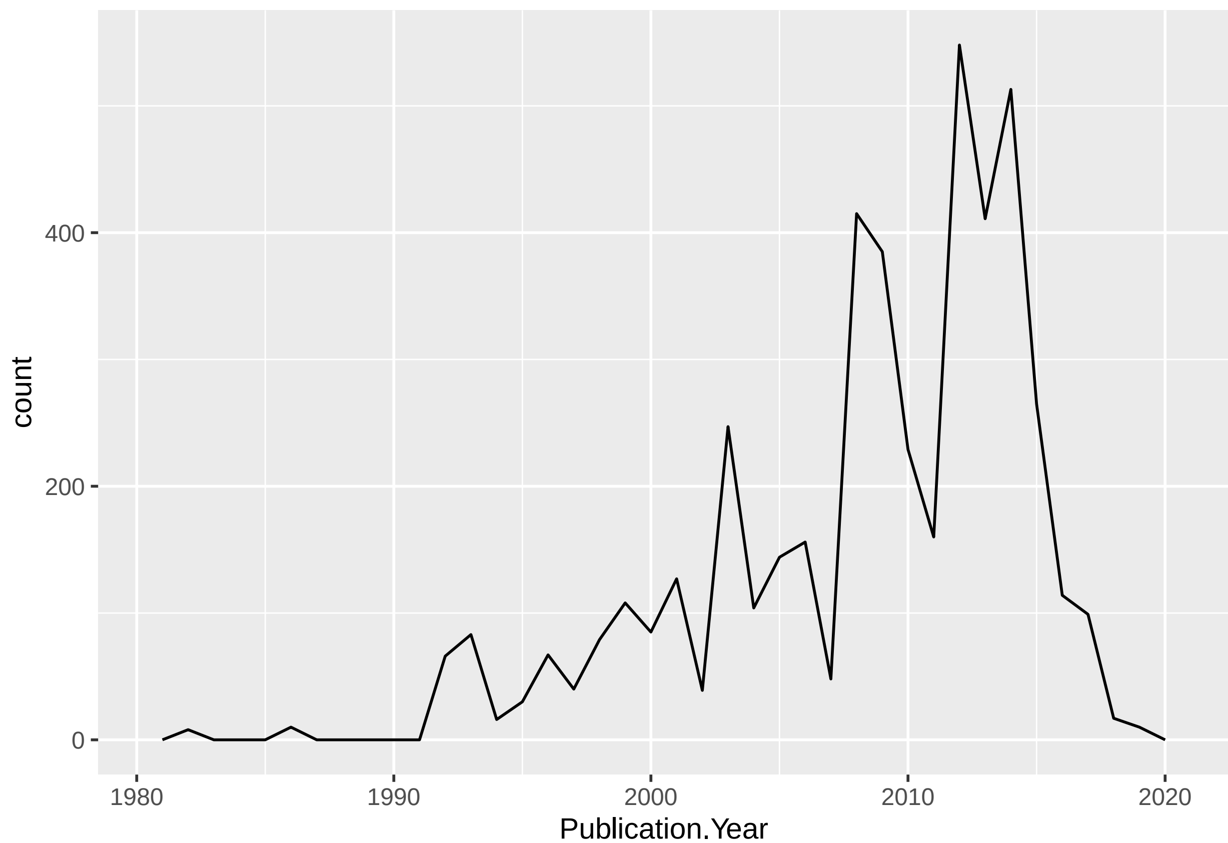
```
## [1] "factor"
```

Answer: some of the values include other symbols like ~ and / and so they cannot be interpreted as numeric values by R

Explore your data graphically (Neonics)

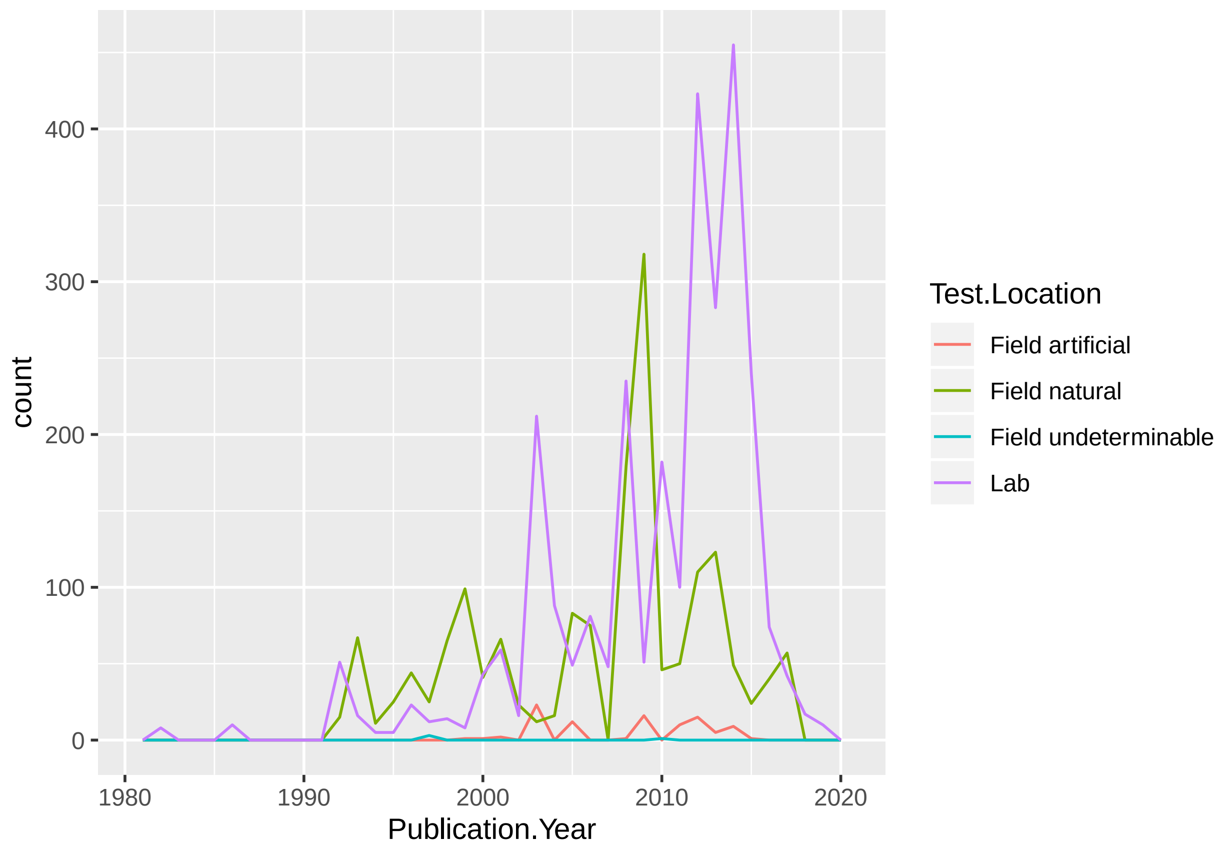
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(neonic.data) +
  geom_freqpoly(aes(x=Publication.Year), binwidth = 1)
```



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

```
ggplot(neonic.data) +
  geom_freqpoly(aes(x=Publication.Year, color = Test.Location), binwidth = 1)
```

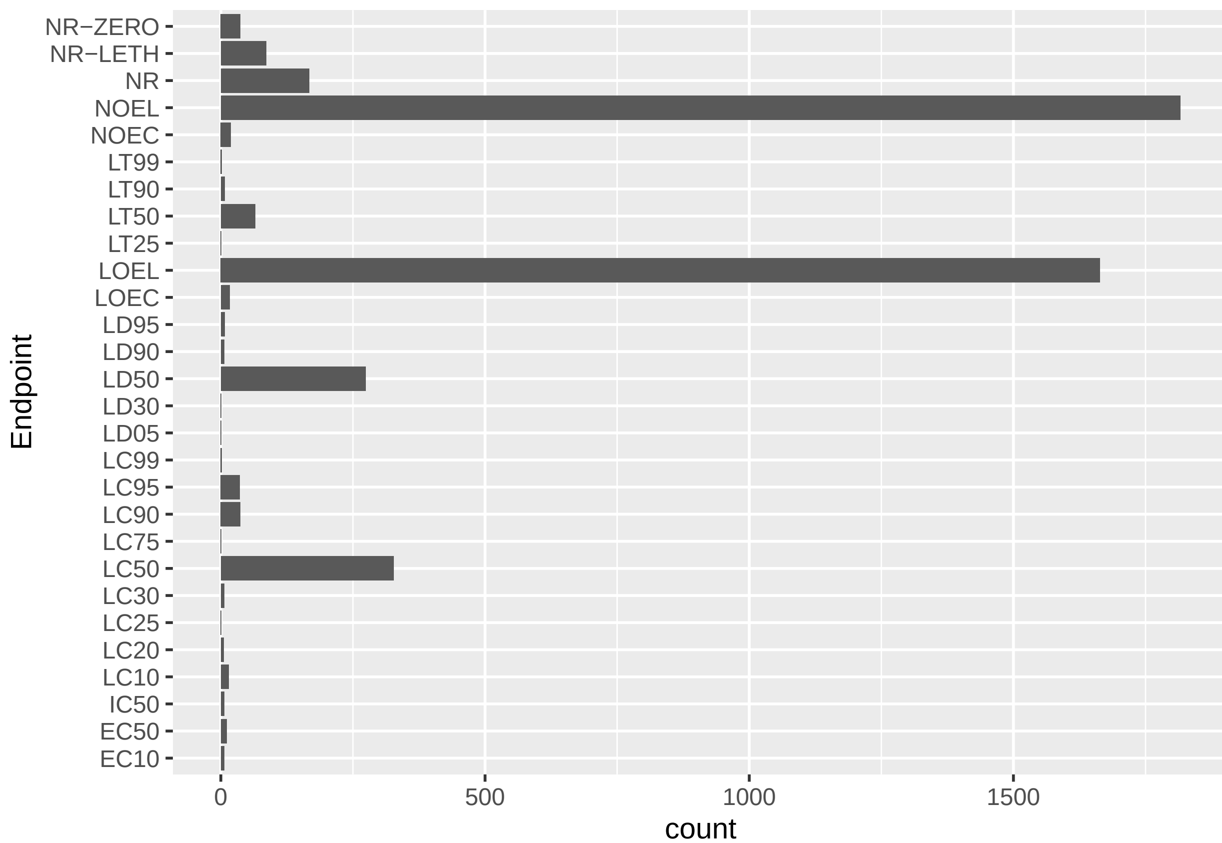


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The number of studies has increased substantially over time. Field studies were more common at first but then lab studies took over in mid 2000s, field studies again had a resurgence but then were quickly replaced with lab studies after 2010. Other test locations remained pretty uncommon

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(neonic.data)+
  geom_bar(aes(x=Endpoint))+
  coord_flip()
```



Answer: most common endpoints reported are NOEL (No observed effect level) and LOEL (lowest observed effect level).

In a study design using multiple different concentrations, the LOEL is defined as the lowest concentration at which a statistically stignificant effect (deviation from the control) is seen. And the NOEL is the highest concentration at which there is no statistically significant difference from the control.

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.

```r
class(litter.data$collectDate)
```

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

```r
litter.data$collectDate <- as.Date(litter.data$collectDate, format = "%Y-%m-%d")
class(litter.data$collectDate)
```

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

```r
unique(litter.data$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`?

```r
unique(litter.data$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
```

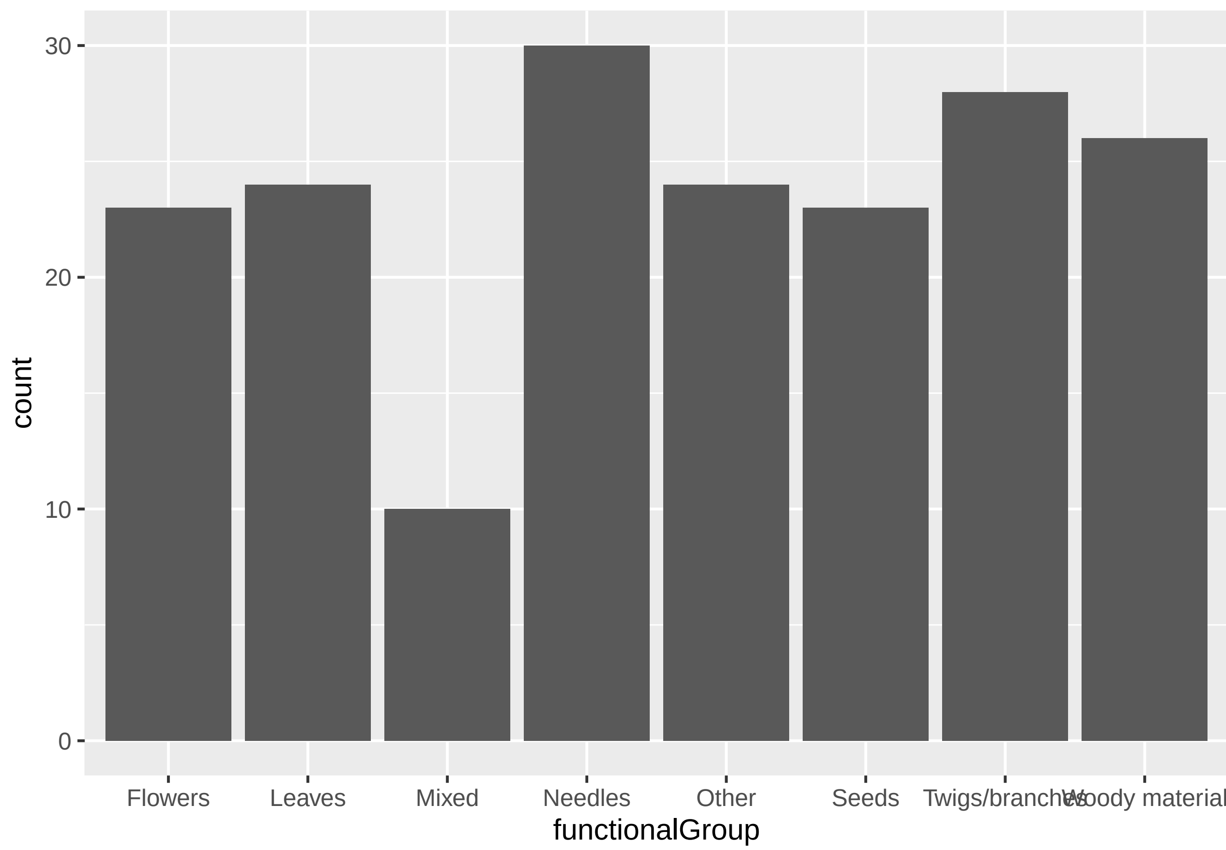
```r
summary(litter.data$plotID)
```

```
## 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. "Unique" lists the different unique values in the column and the total number of levels while "summary" lists the values also with the number of times they occur

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.

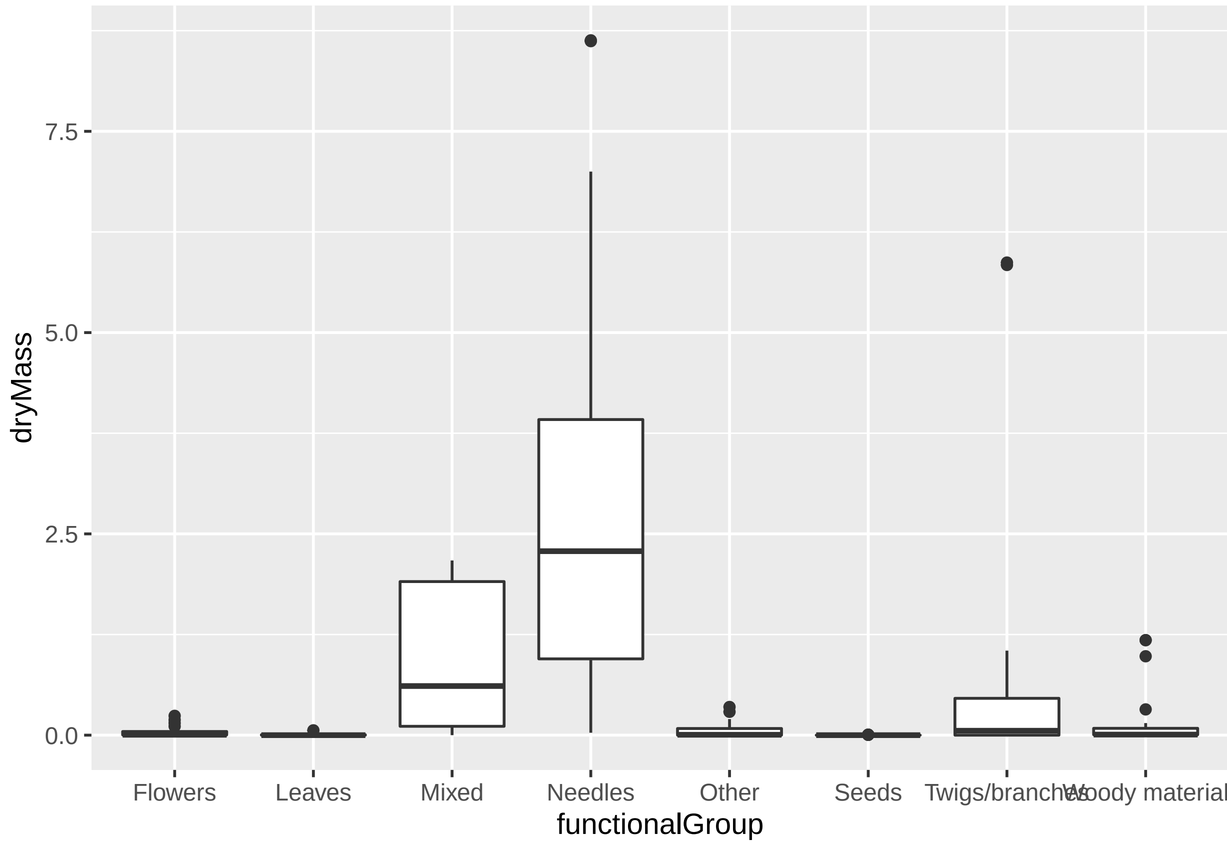
```r
ggplot(litter.data, aes(x=functionalGroup))+
  geom_bar()
```



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

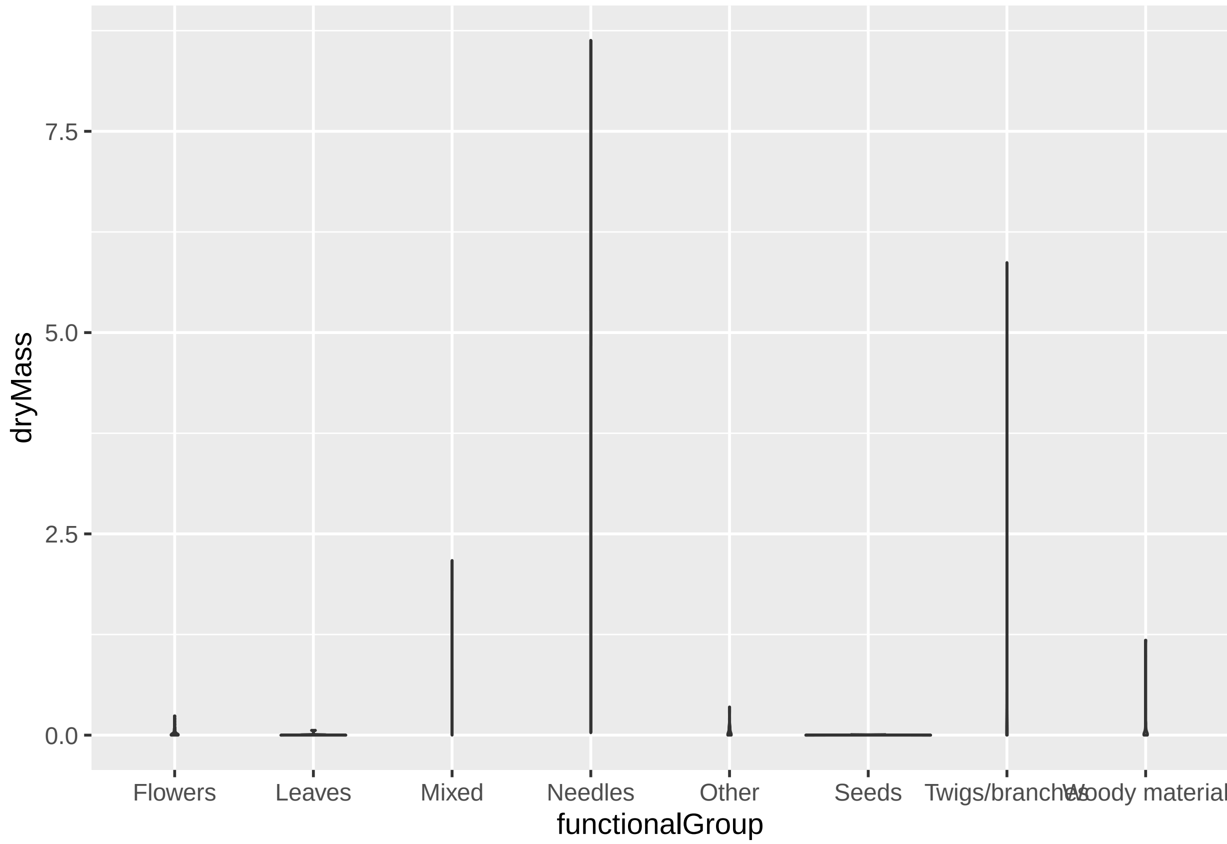
```r
litter.plot <- ggplot(litter.data, aes(x=functionalGroup, y = dryMass))

litter.plot +
  geom_boxplot() #+
```




```
  #scale_y_log10()

litter.plot +
  geom_violin() #+
```




```
#scale_y_log10()
```

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

> Answer: the distributions of dryMass have a wide spread so the violin plot does not effectively show the shape of the distribution unless a log transformation is used

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

> Answer: "Needles" and "Mixed" seem to have the highest dryMass of the different funcitonal groups