# Assignment 3: Data Exploration

Molly Bruce

## 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 <>.

### 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.

```
##Check my working directory
getwd()
```

```
## [1] "C:/Users/mmb88/Desktop/Environmental_Data_Analytics_2021"
```

```
##Set my working directory because, for some reason, it was changing it to the Assignments folder and t
#setwd("C:/Users/mmb88/Desktop/Environmental_Data_Analytics_2021")

##Install & load packages
#install.packages("tidyverse") #installing tidyverse package, commented out for knit
library(tidyverse) #loading the tidyverse package
```

```
## Warning: package 'tidyverse' was built under R version 4.0.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.0.3
```

```
## Warning: package 'tibble' was built under R version 4.0.3
```

```
## Warning: package 'tidyr' was built under R version 4.0.3
```

```
## Warning: package 'readr' was built under R version 4.0.3
```

```
## Warning: package 'purrr' was built under R version 4.0.3

## Warning: package 'dplyr' was built under R version 4.0.3

## Warning: package 'stringr' was built under R version 4.0.3

## Warning: package 'forcats' was built under R version 4.0.3

## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
##Upload datasets
Neonics <- read.csv("C:/Users/mmb88/Desktop/Environmental_Data_Analytics_2021/Data/Raw/ECOTOX_Neonicoti
Litter <- read.csv("C:/Users/mmb88/Desktop/Environmental_Data_Analytics_2021/Data/Raw/NEON_NIWO_Litter_
```

## 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: Neonicotinoids are used to prevent crop/timber decline attributed to particular insects. However, this insecticide can also harm non-target insects. We might be interested in the ecotoxicology of neonicotinoids in order to understand imapcts to non-target insects in the hopes of reducing said impacts.

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: The amount of litter and woody debris can influence stream health. Both litter and woody debris influence sedimentation, turbidity, and erosion for the waterbodies within the forested area being studied. River/water health then influences ecological function more broadly. Furthermore, forested areas are essential carbon sinks and woody debris/litter data can provide insights into the productivity of an area and that productivity's potential implications for its carbon absorption capacity.

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: Researchers rely on paired elevated traps and ground traps to capture litter and woody debris. ~ 4 of these paired traps are placed over a 400 square meter plot. Samples are collected from each trap pair once per year. These collected materials are separated based on their cover-type (needles, leaves, twigs/branches, cones/bark, seeds, etc.) and weighed.

## Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) #this is actually conducting the dimension analysis
```

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

```
#View(Neonics) #this is for my own benefit
#View(Litter) #this is for my own benefit
```

## Check our date column

6. Using the `summary` function on the "Effects" column, determine the most common effects that are studied. Why might these effects specifically be of interest?

```
#The most recent version of R doesn't run traditional summary functions. Instead, users now have to con
EffectsFactors <- as.factor(Neonics$Effect)
summary(EffectsFactors)
```

```
##     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: Scientists are interested in understanding the impacts of insecticides, fungicides, and other chemicals linked to agriculture on insects. Of course, a foremost impact would be insect death at the individual level and insect decline at the population level. Both of these imapcts are observable and of grave concern. However, other impacts may also indicate pernicious impacts to insects caused by neonics. For instance, behavioral changes broadly nad feeding behavior changes more specifically may be a cause for concern. Generally, all of the effects interest scientists because they represent some measure of insect health liked to neonics.

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.

```
#Again, same factor approach as above.
CommonNameFactor <- as.factor(Neonics$Species.Common.Name)
summary(CommonNameFactor)
```

```
##                          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: Excluding the "Other" category, the six most commonly studied species are (1) Honey Bee, (2) Parasitic Wasp, (3) Buff Tailed Bumblebee, (4) Carniolan Honey Been, (5) Bumble Bee, and (6) Italian Honeybee. Each of these insects are polinators. There is particular recognition that insecticides, fungicides, and other chemical applicants to agriculture have particular impact on polinators. These impacts are important because populations of polinators have been declining over the past several decades and these declines pose substantial risks to our methods of food production. Scientists would be interested in studying the impacts of neonics on polinators such as the above six bees in order to understand ways to minimize harms to these polinators linked to agricultural practices.

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] "character"
```

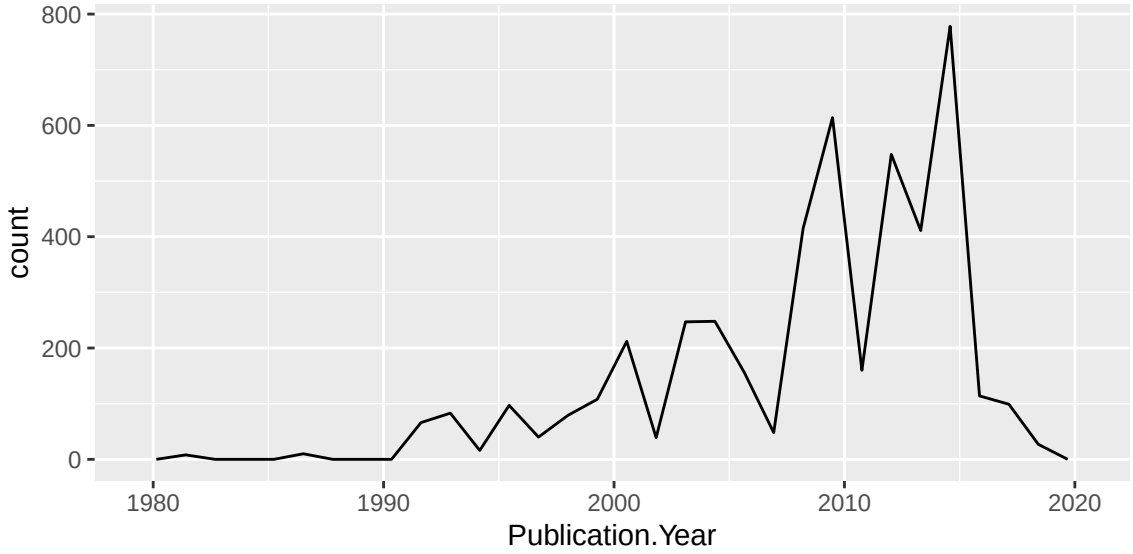
Answer: The class type of Conc.1..Author. is character, likely because it includes some non-numeric inputs such as NR, NR/, and #/. Absent any character-based and symbol-based inputs, RStudio would have read this column as a numeric class.

## Explore your data graphically (Neonics)

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

```
ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year))
```

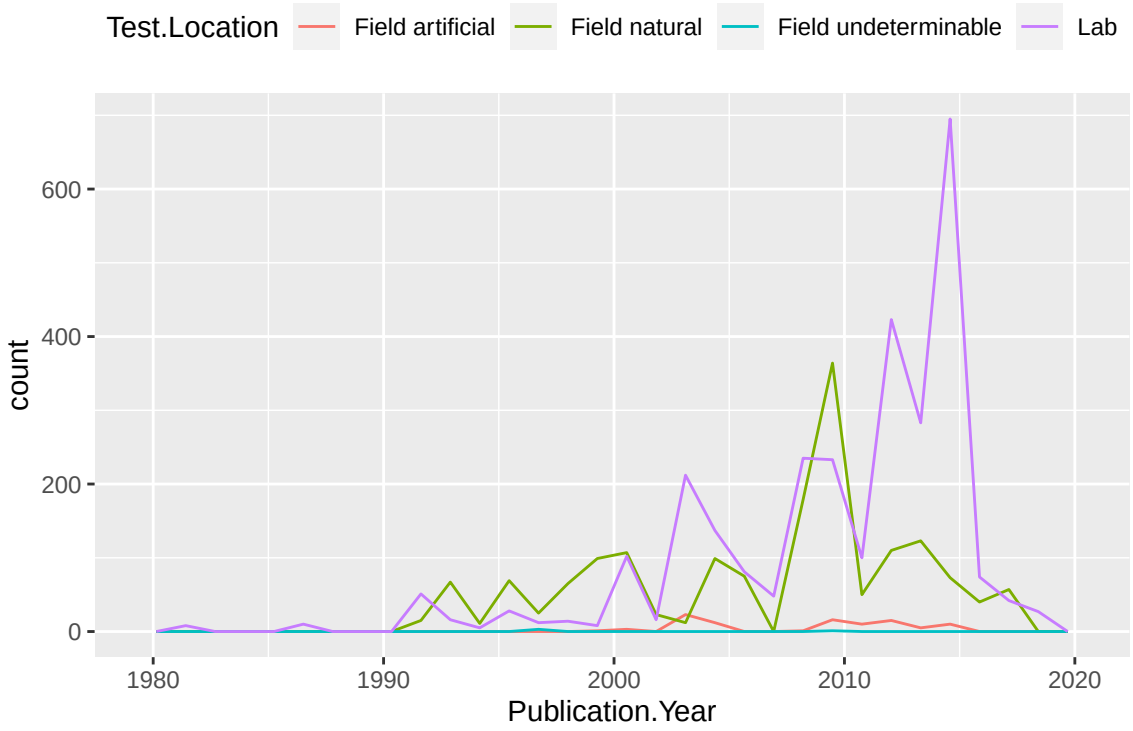
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




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

```
ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year, color = Test.Location)) +
  theme(legend.position = "top")
```

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

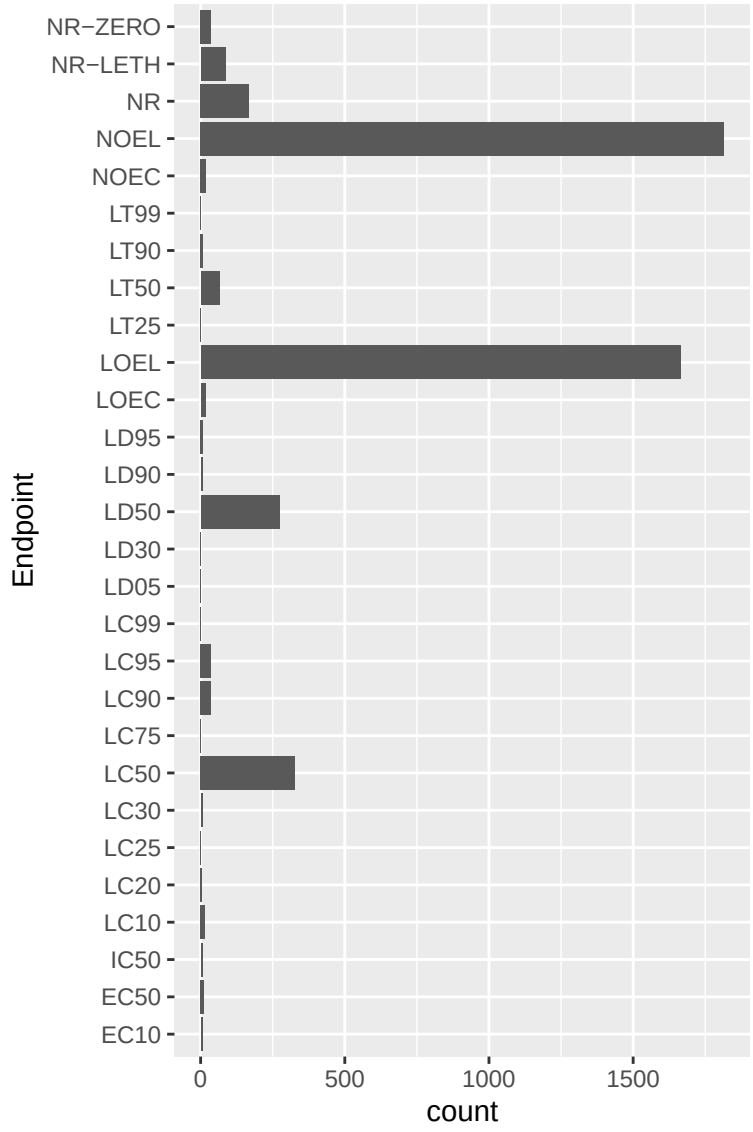



Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: The most common test locations are (1) in the lab and (2) in the natural field. Though natural field-based tests were most common throughout the 90s and around 2009, lab-based tests were far and away the most common test location throughout the 2010s. It's possible that scientists sought more sophisticated testing methods that just weren't possible in the field, thus the transition to more lab-based 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.

```
ggplot(Neonics, aes(y = Endpoint)) + geom_bar() #I chose to put the Endpoint data on the y axis so that
```




Answer: LOEL (lowest observable effect level: the lowest concentration producing effects that were significantly different from response controls) and NOEL (no observable effect level: the highest concentration producing effects that were not significantly different from respons controls) occur most often.

## 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] "character"
```

```
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$plotID)
```

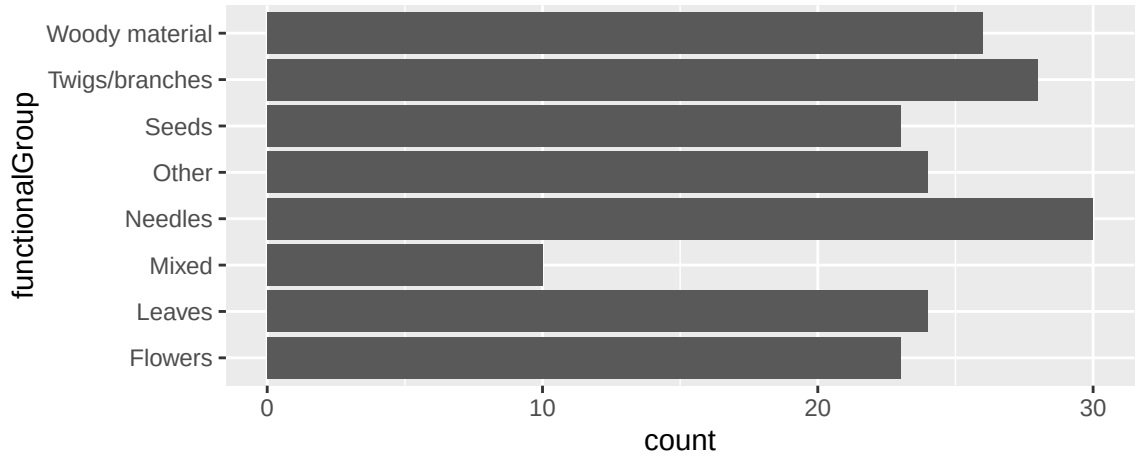
```
##  [1] "NIWO_061" "NIWO_064" "NIWO_067" "NIWO_040" "NIWO_041" "NIWO_063"
##  [7] "NIWO_047" "NIWO_051" "NIWO_058" "NIWO_046" "NIWO_062" "NIWO_057"
```

```
#plotIDFactor <- as.factor(Litter$plotID)
#summary(plotIDFactor)
```

Answer: It looks like 12 plots were sampled. For a character class, the Unique function displays values different from the others, but only displays them once. For a character class, if we convert that class into a factor and then run the summary function (as done above because this version of R requires an additional step for the summary function), we can see not just the different values but also the count for each of those different values.

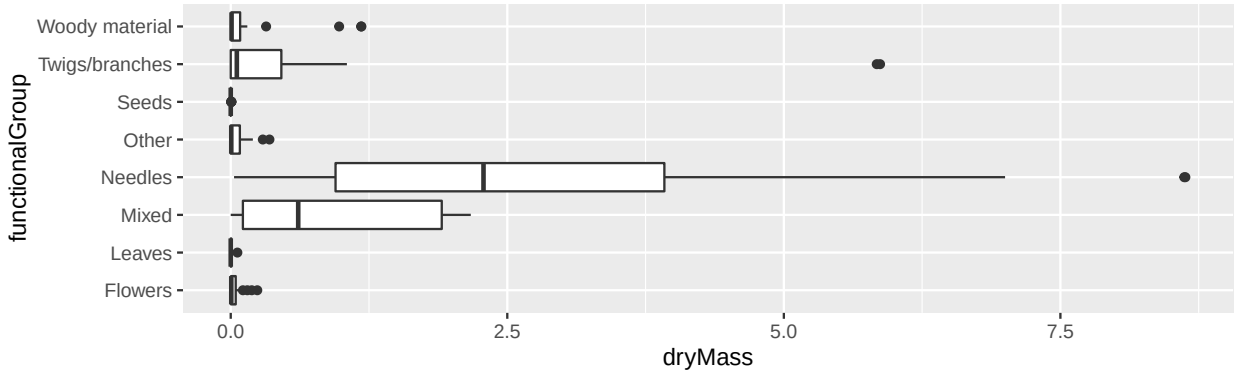
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(y = functionalGroup)) + geom_bar() #again, I chose to put the functionalGroup on the
```

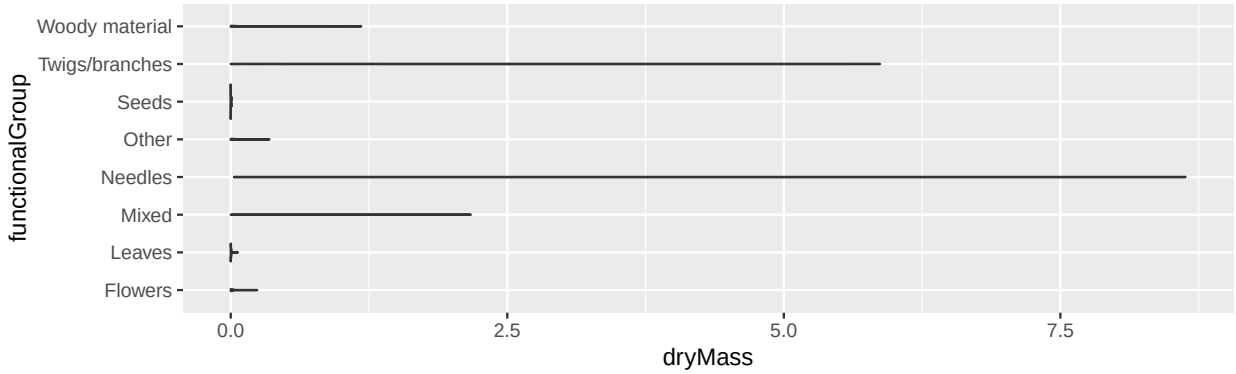



15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
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: The boxplot does a more effective job of displaying this type of spread-out data because it shows the range of values and the distribution of values within that range using lines and boxes. It does a better job of displaying outliers and quartiles. Alternatively, if our data were clustered very heavily around particular measurements (dryMasses) with only a couple outliers, the violin plot might be a more useful tool. However, because our data is more spread out and we have wide variability in dryMass depending on the functionalGroup, the box plot's visualization captures the data better.

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

> Answer: Needles and mixed content tend to have the highest biomass. This makes sense because needles are more dense and take longer to biodegrade than leaves, powers, etc. Needles are also more abundant at higher altitudes such as Niwot Ridge which sits at ~10,000 ft ASL.