# Assignment 3: Data Exploration

Andrea Gonzalez Natera

## 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.

```
#Setting working directory to Environmental_Data_Analytics_2021
getwd()
```

```
## [1] "C:/Users/andre/Documents/NSOE-MEM 2019-2021/Spring 2021/Data Analytics/Environmental_Data_Analy
```

```
#Packages
#install.packages("dplyr")
#install.packages("ggplot2")
#install.packages("tidyverse")

library(dplyr)
library(ggplot2)
library(tidyverse)
library(lubridate)
```

```
#Uploading datasets
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)
```

### 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: Th ecotoxicology of neonicotinoids in insects is something we want to study because we need to understand and quantify the effectivness of neonicotinoids in both pests such as fleas and beetles and also on beneficial insects such as bees

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: Because litter and woody debris will have impact on the belowground biomass of a forest. They are also fuel sources and can be indicators of the scale and probability of a forest fire.

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: Woody debirs is collected in ground traps that are then sampled every year. Litter and fine woody debirs are collected in elevated traps that are selected randomly within the 90% flux footprint of the primary and secondary airsheds.Elevated traps are sampled once every two weeks in decidous forest sites and once every 1-2 months in evergreen sites.

## 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 on the "Effects" 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: Population is the most commonly studied effect, followed by mortality. This is probably due to environmental studys on the effects of neonicotinoids in beneficial insect populations and mortality to evaluate the environmnetal impact.

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: Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, Italian Honeybee.

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"
```

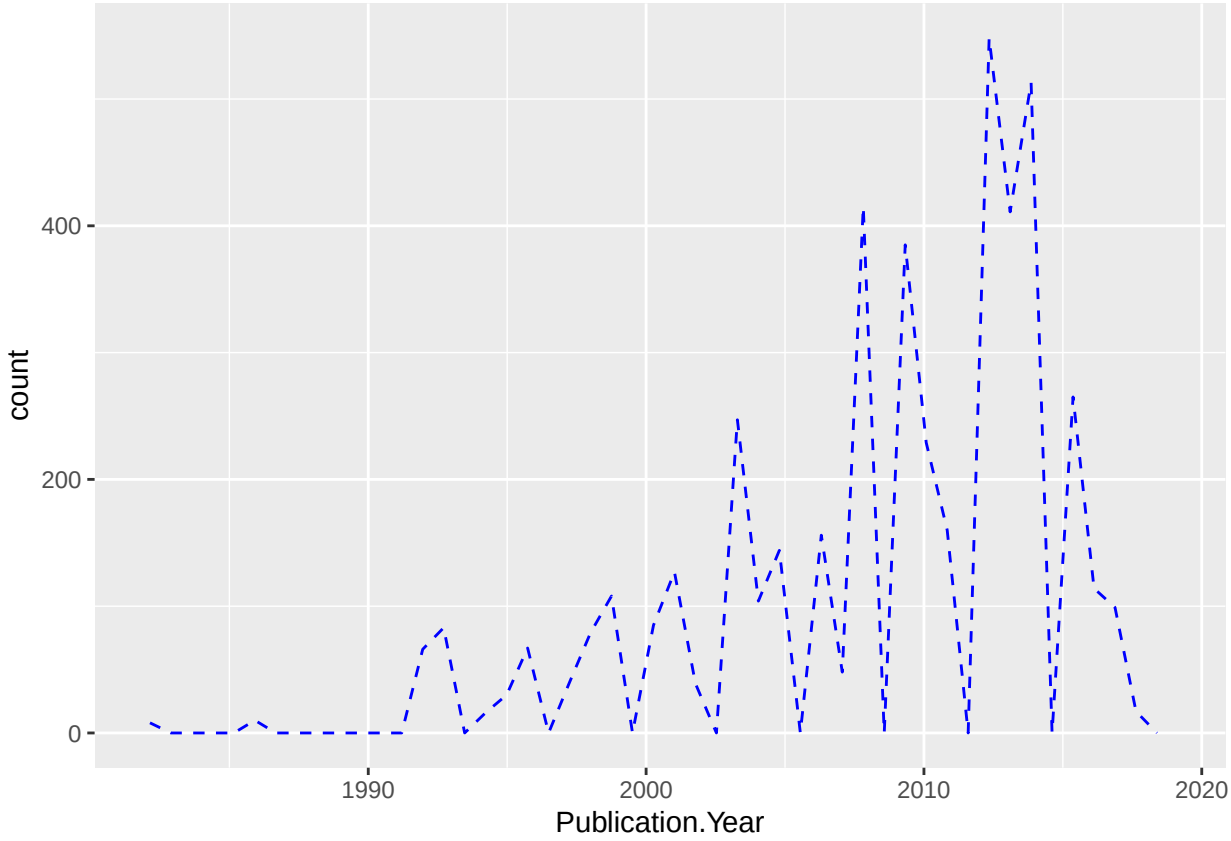
Answer: Conc.1..Author is a factor, it is not numeric because although it is a concentration, measurments are taken in different unit for different treatments for example Al mg/L or fl oz/acre. This means that mathematically they are not comparable and therefor it makes more sense to think of the data as categorical rather than numeric.

## Explore your data graphically (Neonics)

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

```
#It is in integer, so I will change that to date
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50, color = "blue", lty = 2) +
  scale_x_continuous(limits = c(1982,2019))
```

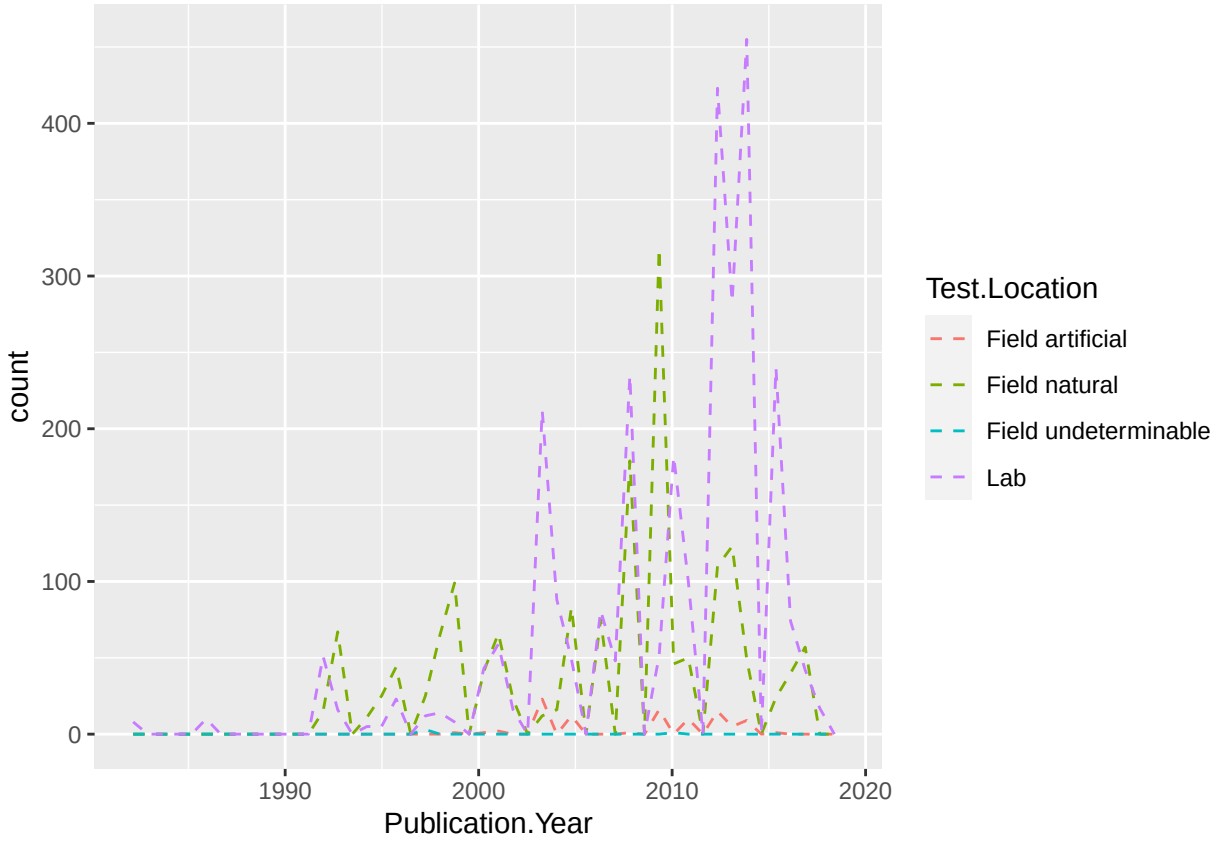
```
## Warning: Removed 3 row(s) containing missing values (geom_path).
```




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 = 50, lty = 2) +
  scale_x_continuous(limits = c(1982,2019))
```

```
## Warning: Removed 12 row(s) containing missing values (geom_path).
```

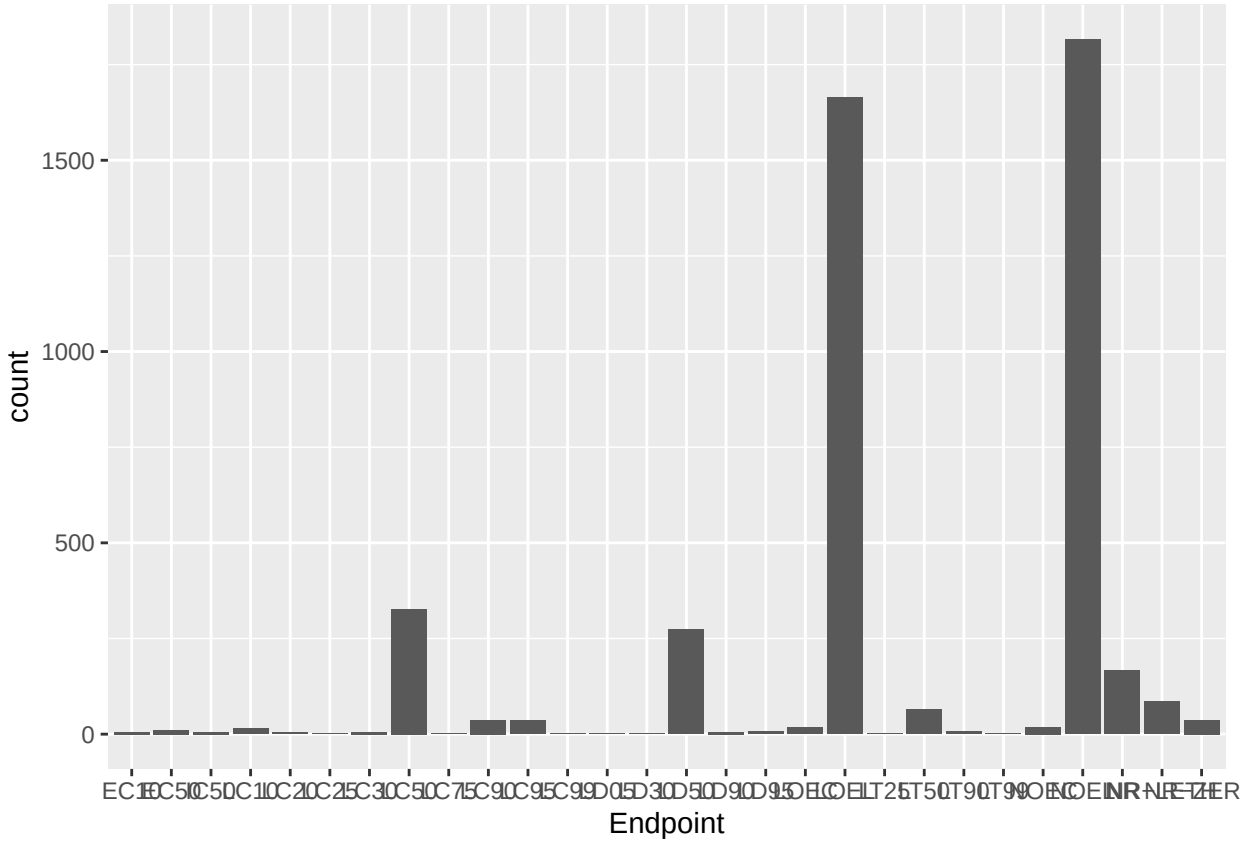



Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations since 2000 are Labs. Before 200 the msot common locations where Field natural, they still remain the second monst common location and like labs their frequency has increased over time.

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()
```




Answer: The two most common Endpoints by far are NOEL and LOEL. NOEL is defined as No-observable-effect-level. This means that even at the highest experimental dose there was no significant difference form the control group in the statistical test. LOEL is defined as Lowest-observable-effect-level. This means that at the lowest experimieltal dose there were significantly different results than those fom the control group in the statistical test.

## 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"
```

```
#collectDate is a factor
Litter$collectDate <- ymd(Litter$collectDate)
class(Litter$collectDate)
```

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

```
#CollectDta is now a Date
unique(Litter$collectDate)
```

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

```
#Litter was sampled during on August 2 and 30 of 2018
```

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`?

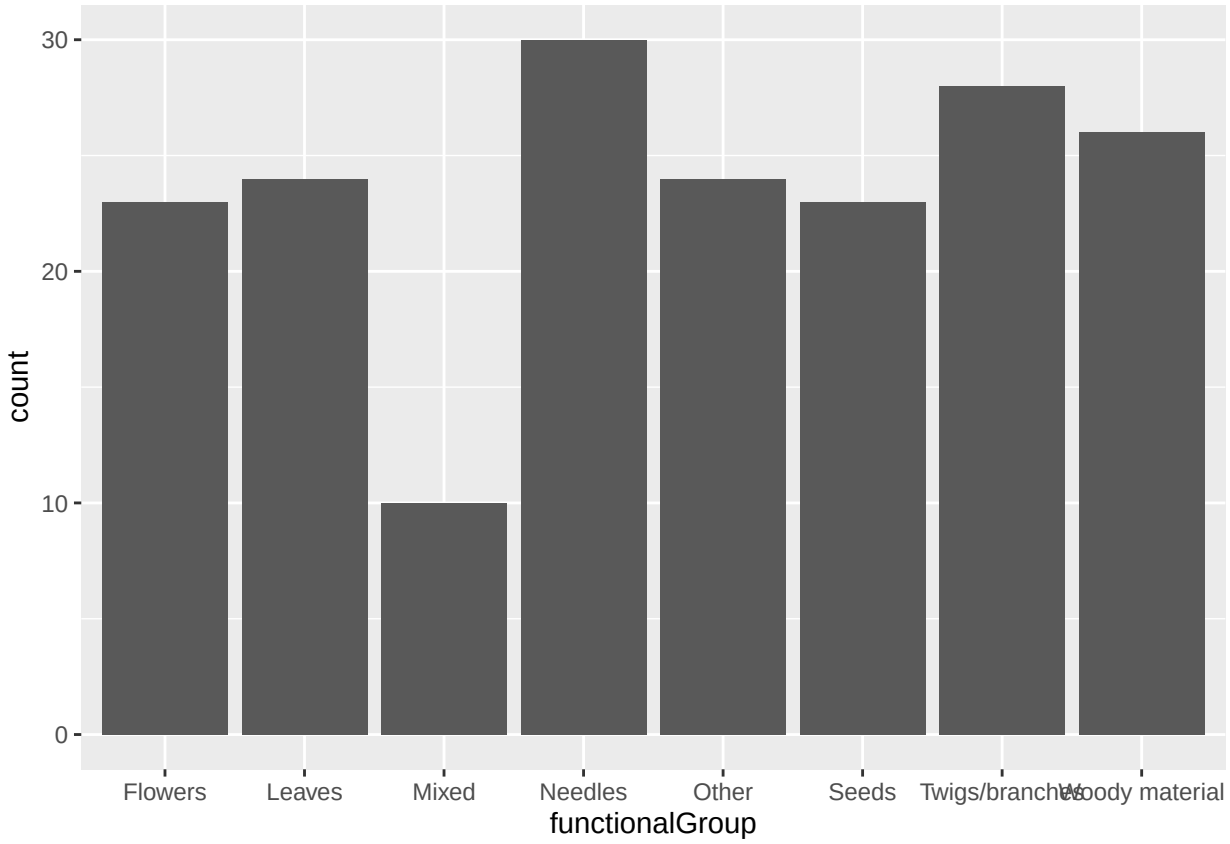
```
unique(Litter$plotID)
```

```
##  [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: There were 12 plots sampled in Niwot Ridge. Using the 'unique' function we very quickly get the number of unique factors (levels) wihtout having to manually count. The summary function gives us the number of records for each plot but we have to manually count each plot to get the number of unique values.

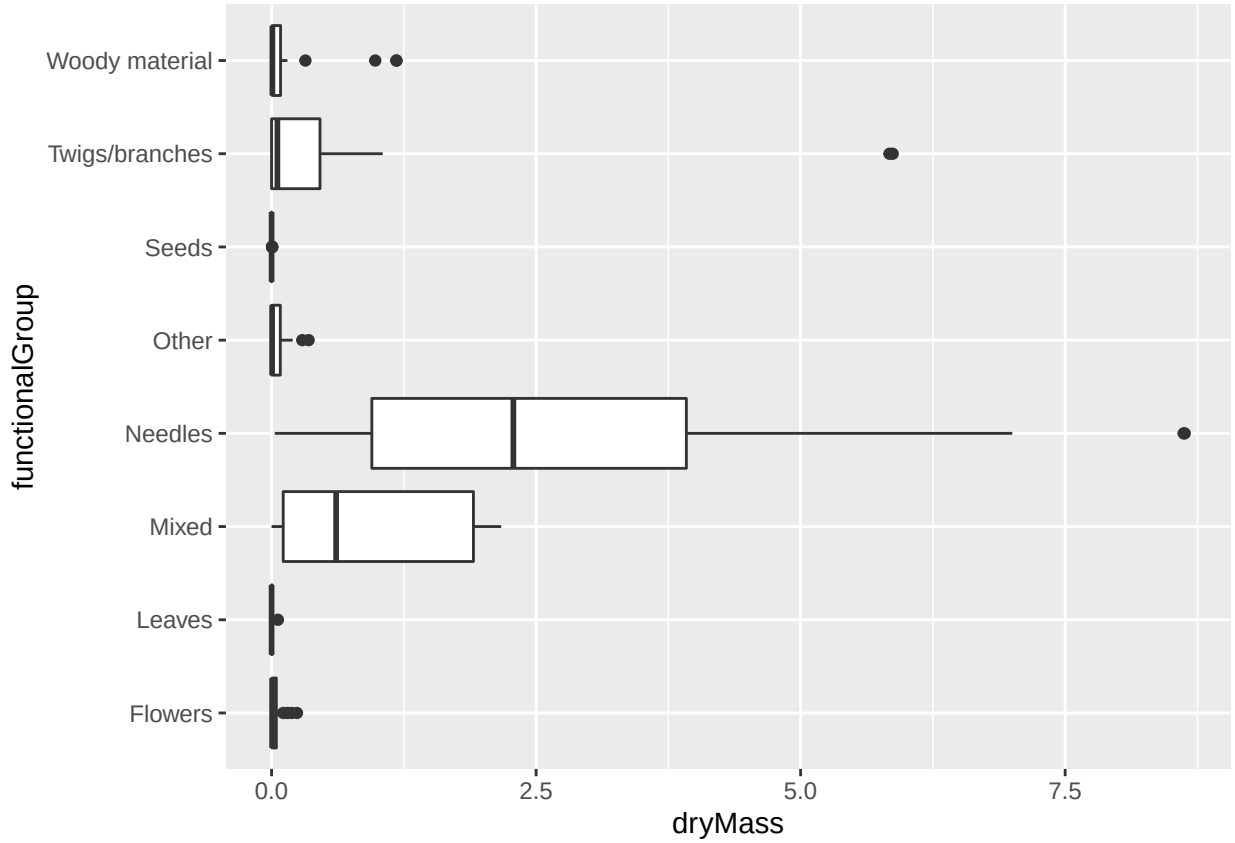
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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



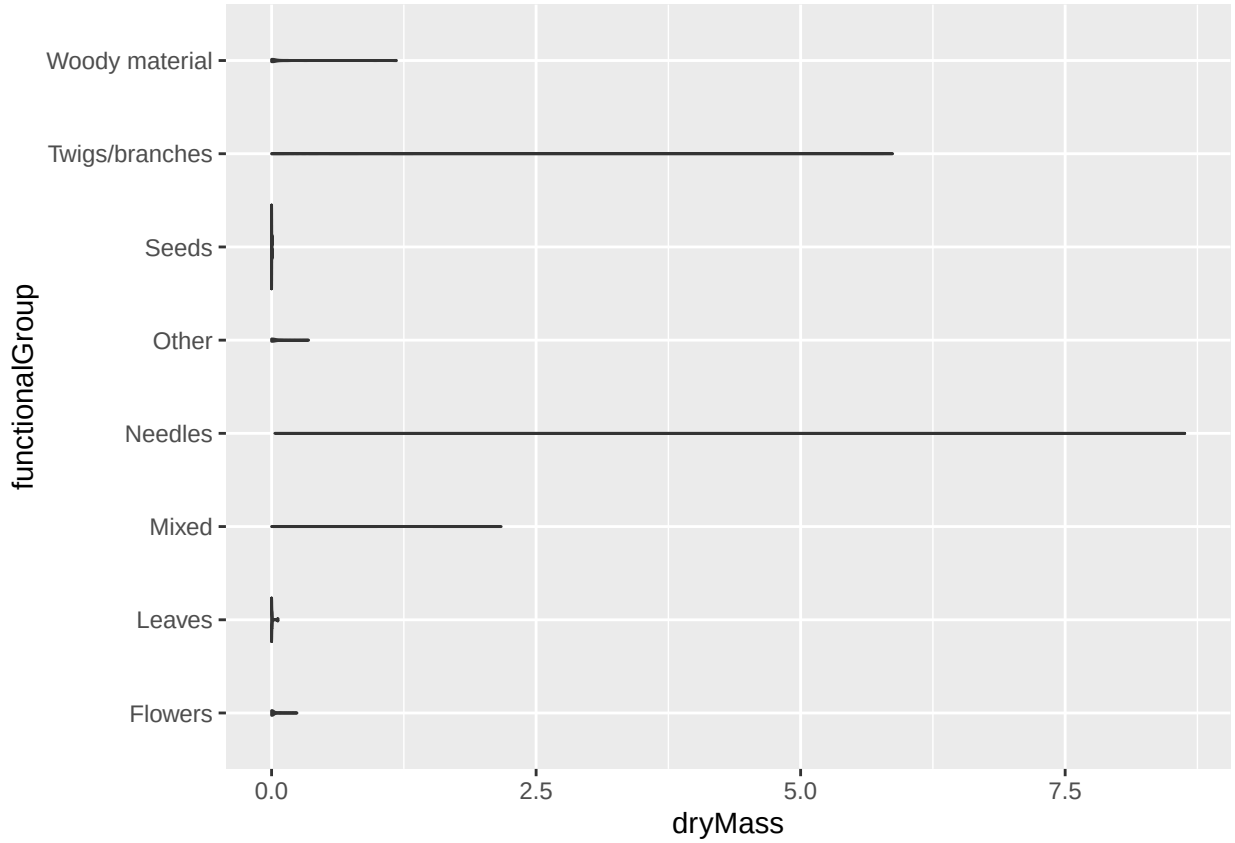

15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




```
#Violin Plot

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: Because the boxplot provides more information regarding the distribution of the data like the mean, quartiles and outliers.

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

Answer: Needles have tend to have highest biomass