# Assignment 3: Data Exploration

Austin Guimond

## 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.

```
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = TRUE)
#head(Neontics)
Litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv",stringsAsFactors = TRUE)
#head(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 ecotoxicology of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: The ecotox data has implications for both human and environmental health. The neontics data shows what species are targeted by specific insecticides and what plant species they are applied to. Using this information we can make conclusions about where the insecticide

have been applied and what concentrations are appplied for treatment. This infomation could be used to cross reference insecticide hotspots and determine what agricultural products likely contain traces of certain pesticides. If there are human health trends associated with agricultural products, we coud use this information to see if adverse health effects could be the result of the pesticide.

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 litter that falls to the ground at these test sites are likely analyzed to evaluate forest health and productivity. Healthy forests will have lots of seeds and flowers that drop to the ground as a sign that the plant community was healthy enough to reproduce. On the other hand, large loss of needles, branches, or leaves could be signs of pest infestations. In Colorado, pine beetles kill lodgepole pines resulting in loss of needles. The Niwot Ridge data could also be tracking pest infestations like the pine beetle.

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. Sampling occurs in 20 40x40 plots. 2. One elevated and one ground trap are deployed every 400 square meters. 3. Ground sampling traps are collected once a year while elevated traps are collected at a rate dependent on the species compostion. For example, elevated traps in evergreen forests are measured once every 1-2 months.

## Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#find information and dimensions about Litter data set
#str(Litter)
dim(Litter)
```

```
## [1] 188  19
```

```
length(Litter)
```

```
## [1] 19
```

```
#find information and dimensions about Ecotox data set
#str(Neonics)
dim(Neonics)
```

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

```
length(Neonics)
```

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

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?

```
#summarize effects of Ecotox data
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 common effects are Mortality (1493), Population (1803), and Behavior (360). All of these effects are related to environmental impacts of pesticide applications on target species. One of the most common effects listed is mortality which is being measured as a direct result of the pesticides. The other two common variables are population and behavior which are measuring how the populations and behaviors of the target species are being influenced by specific pesticide applications. These charcters could also be used to study adverse environmental impacts of pesticide applications.

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.

```
#summarize species in ecotox data
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: 1. Honey Bee 2. Parasitic Wasp 3. Buff Tailed Bumblebee 4. Carniolan Honey Bee 5. Buble Bee 6. Italian Honeybee All of these species are pollinators including the parasitic wasp. Pesticide data would be especially important for understanding adverse reactions to pollinators including morbitity or changes in behavior. If pesticides that were applied to kill a pest had adverse effects for pollinators, the neonics dataset could be used to look at the scale of the adverse effects.

8. Concentrations are always a numeric value. What is the class of Conc.1..Author. in the dataset, and why is it not numeric?

```
#find class
class(Neonics$Conc.1..Author.)
```

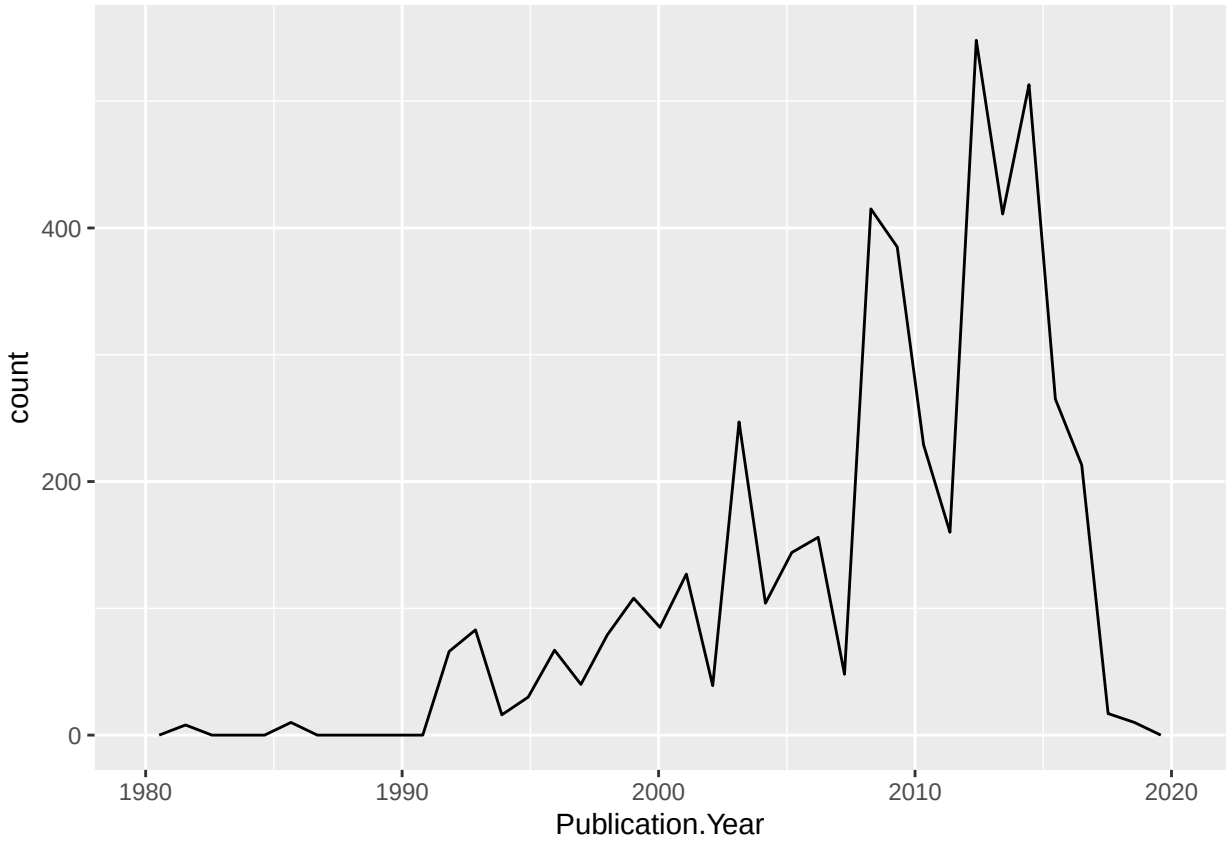
```
## [1] "factor"
```

Answer: The Conc.1.Author is factor data because it contains non-numeric values including / to indicate per unit measurment and NR. If the column was numeric value none of the non-numeric characters would register in the dataset. Numeric data can only contain numeric values while factor data can store a select amount of non-numeric values.

## Explore your data graphically (Neonics)

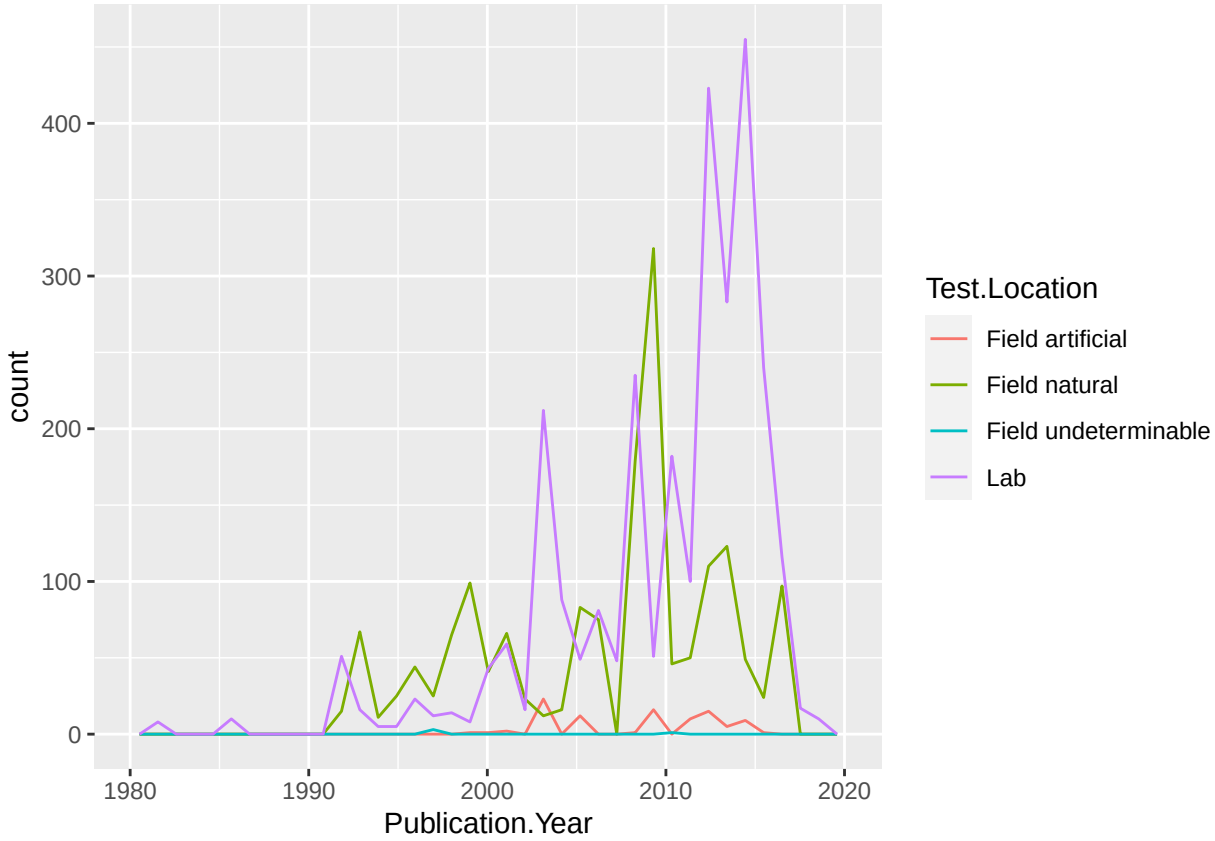
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
library(ggplot2)
#plot publications per year
ggplot(Neonics) + geom_freqpoly(aes(x= 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.

```
#plot publications by test locations
ggplot(Neonics) +
  geom_freqpoly(aes(x= Publication.Year, color = Test.Location), bins =37)
```

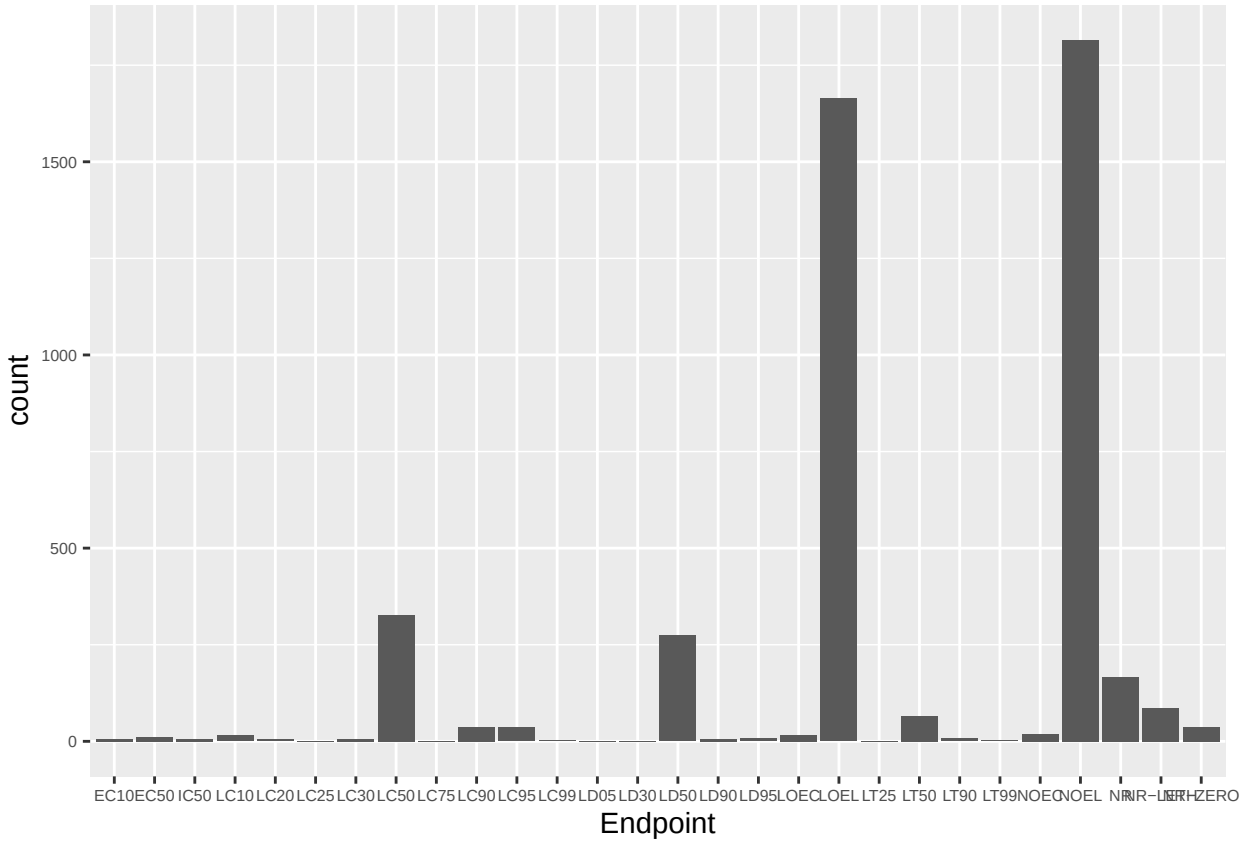



Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations especially from 2010 to 2020 are labs. From 1990 to 2000 field natural test locations were the most common as well as from 2008-2010. The lowest rates of test locations were field artificial and field undeterminable.

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.

```
#plot bar graph of data endpoints
ggplot(Neonics, aes(x=Endpoint)) +
  geom_bar() + theme(axis.text = element_text(size = 6))
```




Answer: The two most common endpoints are NOEL and and LOEL. NOEL is terrestrial data where pesticide application had no observable effects. LOEL is also terrestrial data but the lowest possible dosage produced significantly observable results.

## 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.

```
#find collection date class
class(Litter$collectDate)
```

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

```
#convert date as factor to date
Litter$collectDate <- as.Date(Litter$collectDate)
#find litter collection dates
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`?

```
#find number of plots in Niwot study
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
```

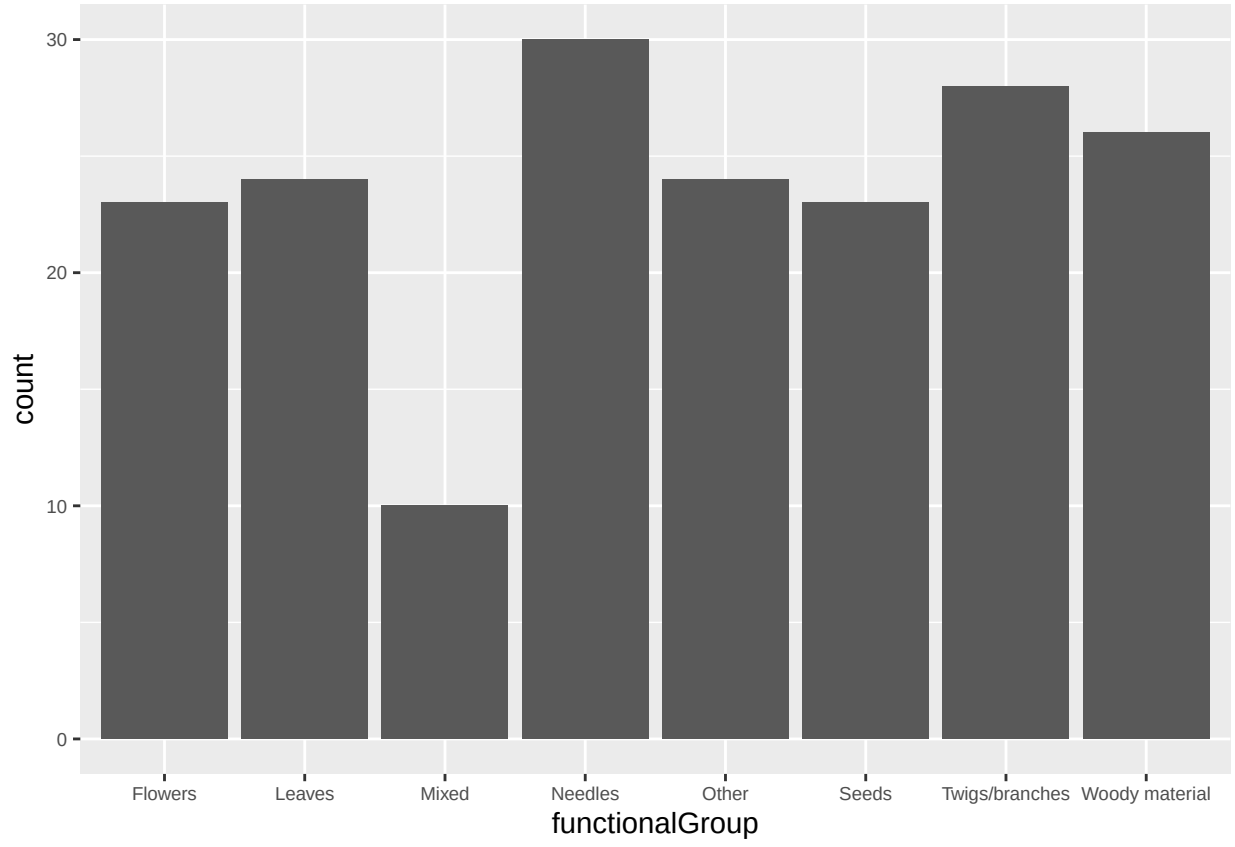
```
summary(Litter$plotID)
```

```
## 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: There were twelve plots sampled at Niwot Ridge. The unique function returns the different plots listed incrementally by the amount of occurances. The summary function returns the different plots that were collected on Niwot Ridge and how many data points for each plot are listed in the column.

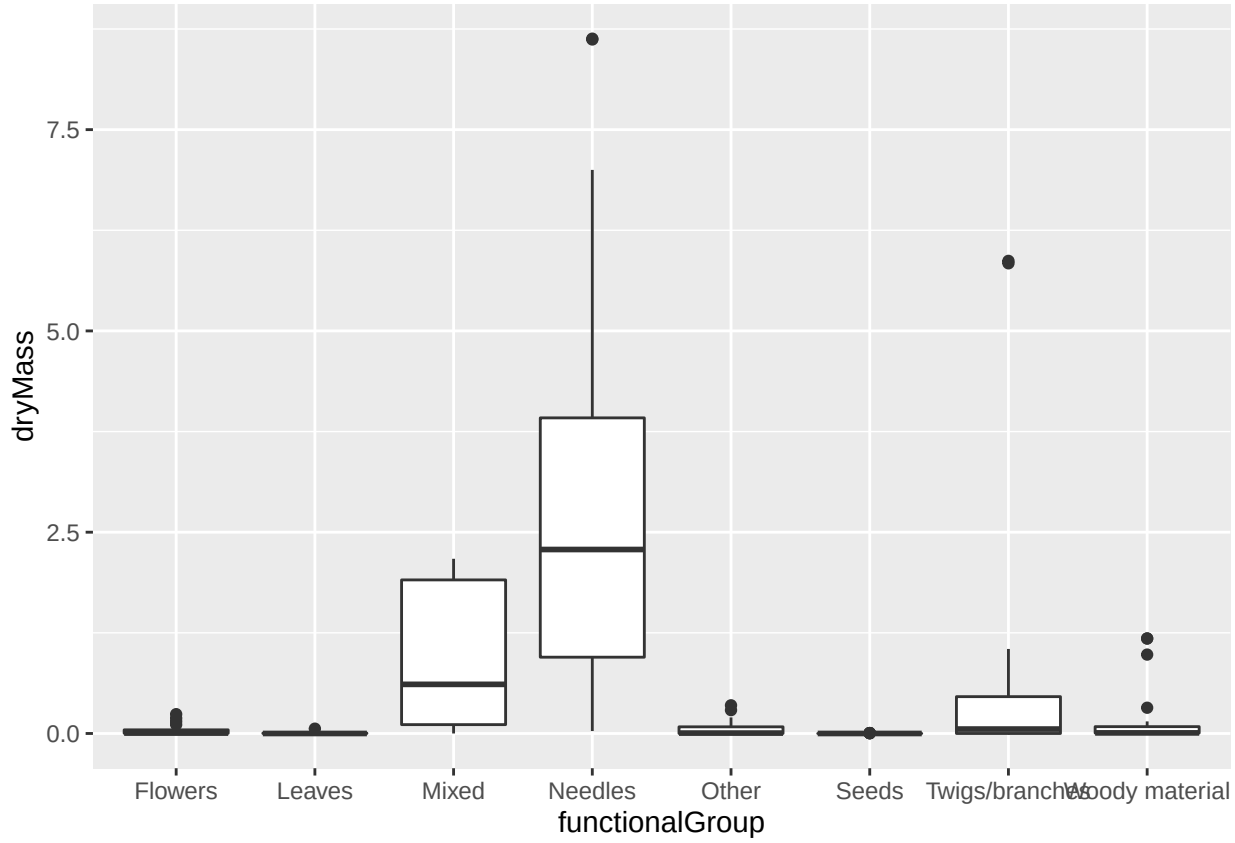
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#graph litter types
ggplot(Litter, aes(x=functionalGroup)) +
  geom_bar() + theme(axis.text = element_text(size = 7))
```



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

```
#create boxplot of drymass measures by functional group
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass))
```

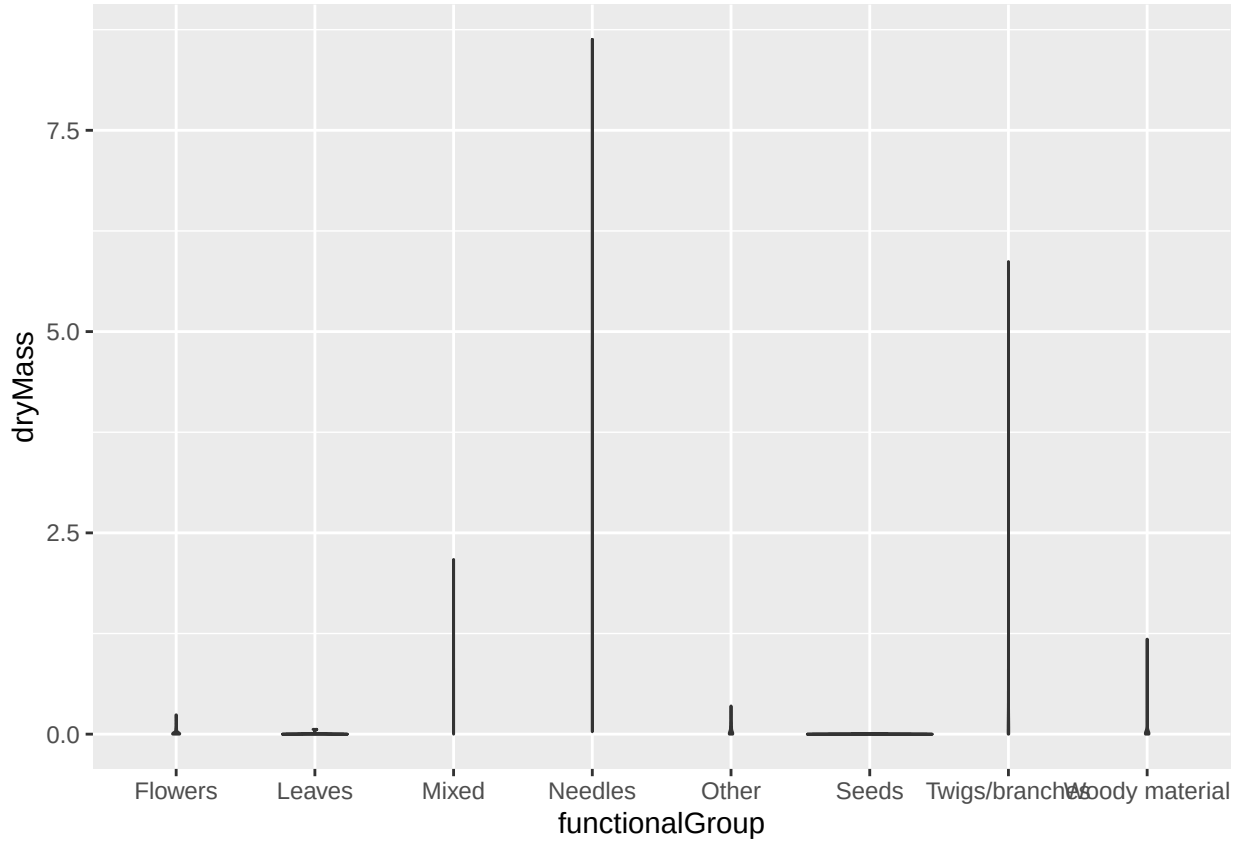



```
#create violin plot of drymass by functional group
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75))
```

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

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

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




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

> Answer: The distribution of data is fairly even across the datasets and there were very few clusters within the dryMass data. Because there are no clusters, the violin plots do not eleongate outwards and just appear as elongated lines. Violin plots are best for data sets where there are clusters of data. In this case the most informative information is the mean and quantiles which are displayed more clearly in the boxplot.

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

> Answer: The litter that has the highest biomass at these sites is pine needles and mixed drymass. This makes sense because Niwot Ridge is in the Rocky Mountains of Colorado where pine trees dominate the landscape.