Assignment 3: Data Exploration

Katie Krejsa

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.

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

```
## [1] "/Users/kaitlynkrejsa/Desktop/Duke/Classes/ENVIRON 872 - Environmental Data Analytics/Environment
```

```
# loading packages
library(tidyverse)

# 3. Importing datasets
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 the ecotoxicology of neonicotinoids on insects because different insecticides are more or less effective depending on a number of variables such as the species of insect, the life stage of the insect, the age of the insect, application time of year, number of chemical applications, and concentrations of chemical applications, among others.

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: We might be interested in studying litter and woody debris that falls on the ground in forests because it can provide information on decomposition rate and soil development. It can also provide information on how much carbon is being store in those sinks.

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: * "Mass data for each collection event are measured separately for the following functional groups, to an accuracy of 0.01 grams." * "One litter trap pair (one elevated trap and one ground trap) is deployed for every 400 m2 plot area, resulting in 1-4 trap pairs per plot." * "Ground traps are sampled once per year. Target sampling frequency for elevated traps varies by vegetation present at the site, with frequent sampling (1x every 2weeks) in deciduous forest sites during senescence, and in- frequent year-round sampling (1x every 1-2 months) at evergreen sites."

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset? 4623 rows and 30 columns

```
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?

```
# Effect is a character, so cannot run a summary on it. Need to convert to a factor first

Neonics$Effect <- as.factor(Neonics$Effect)
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 studied are population, mortality, and behavior. These effects are specifically of interest to determine how effective the insecticide is on reducing the population size of the target insect (population), on killing the target insect (mortality), and on behavioral changes of the target insect, such as feeding on the target crop (behavior). These effects are also of interest regarding non-target insects, such as pollinator species, because it may be important to make sure the insecticide is not reducing their population, not killing them, and not changing their pollinating behaviors.

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.

```r
# species.common.name is a character, so cannot run a summary on it. Need to convert to a factor first

Neonics$Species.Common.Name <- as.factor(Neonics$Species.Common.Name)
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: The six most commonly studied species in the dataset are: Other, Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, and Bumble Bee. All of these species are pollinator species. Pollinator species are critical for maintaining high crop yields. It is important to study the effects of insecticides on pollinators to make sure the insecticides are not having a deleterious effect on these species.

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

```r
class(Neonics$Conc.1..Author.)
```

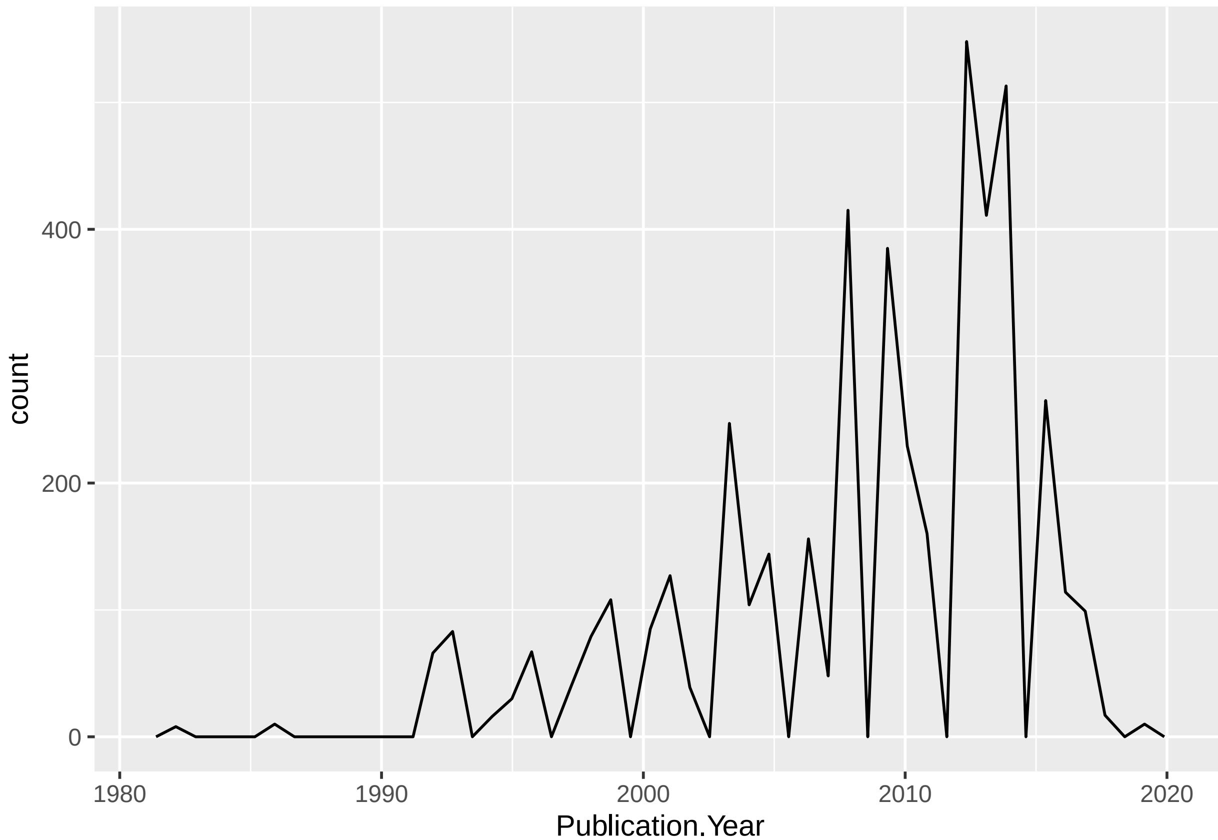
```
## [1] "character"
```

Answer: It is a character. It is not numeric because it has some non-numeric values like "NR" and "~10", so the class of the entire column defaults to character.

Explore your data graphically (Neonics)

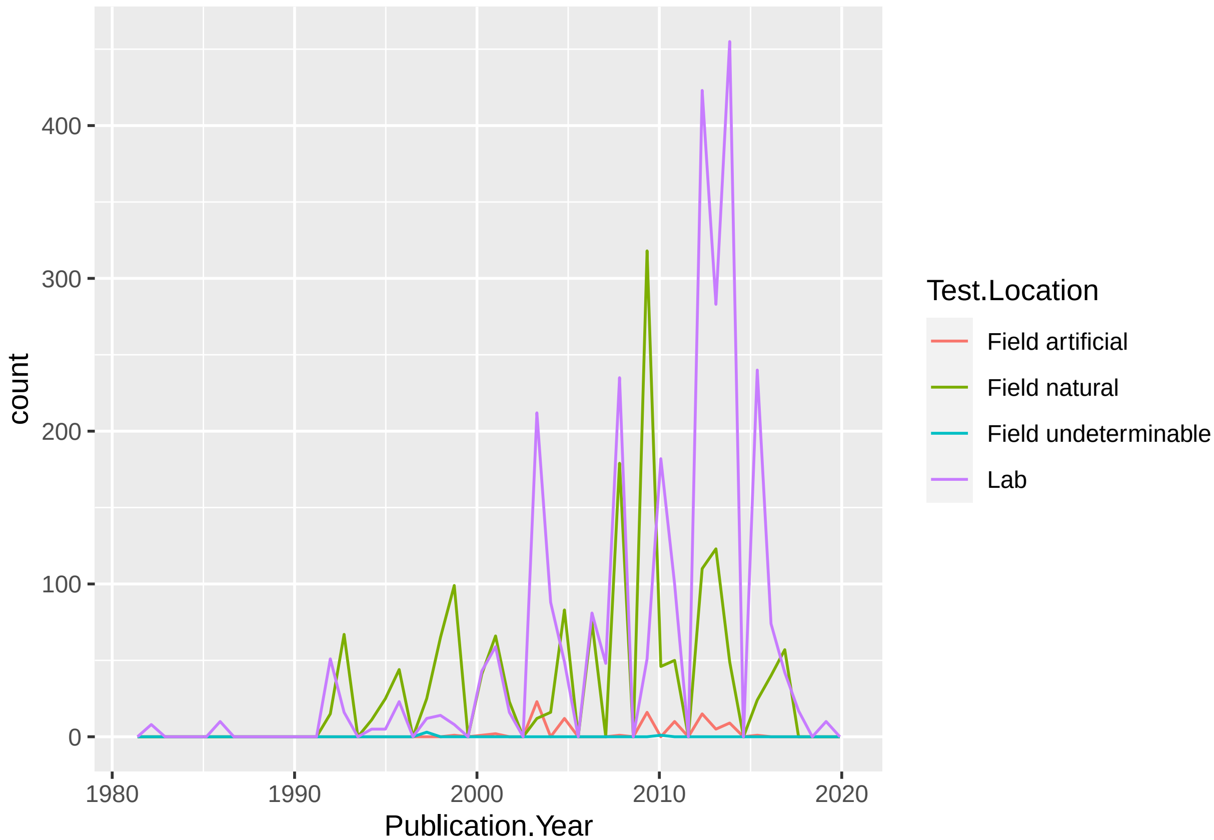
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```r
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50)
```



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)
```

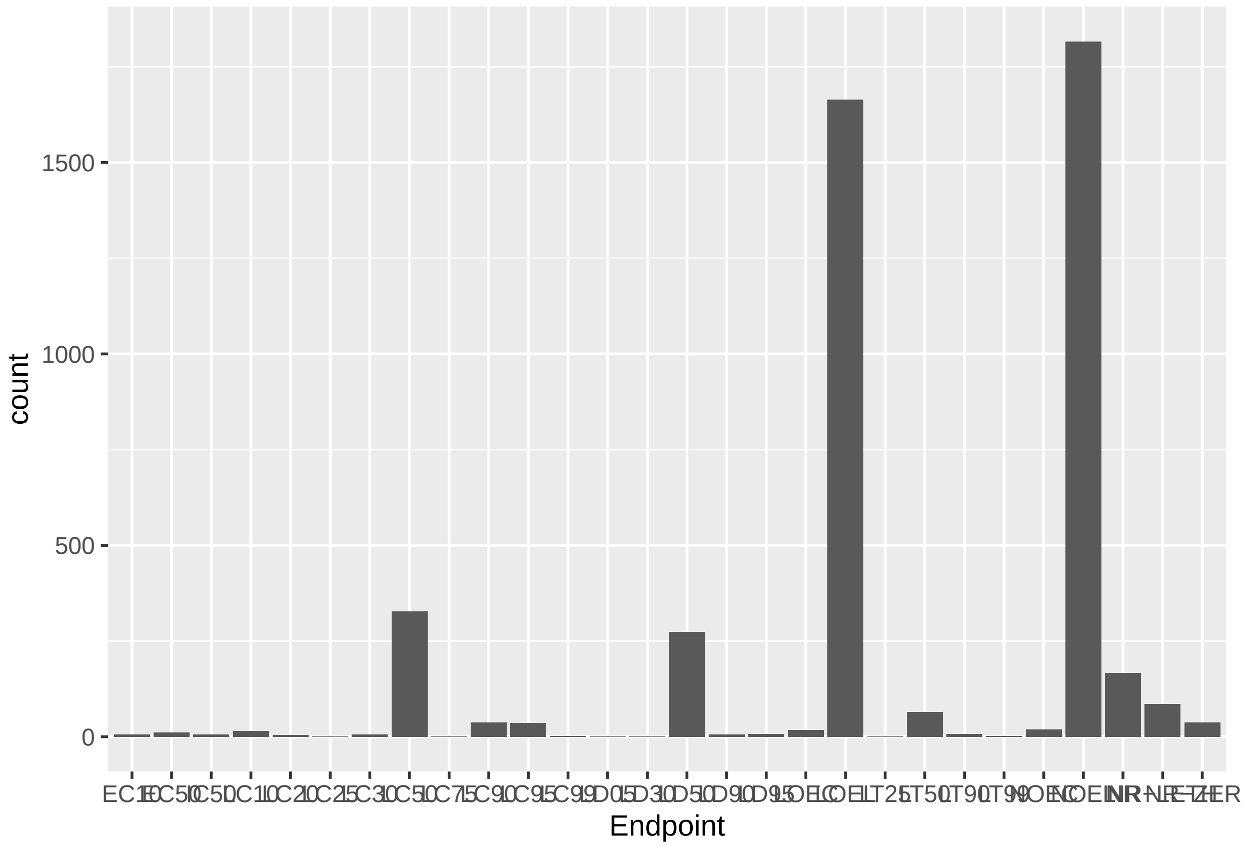


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are lab and field natural. Lab and field natural locations differ quite a bit overtime, but overall these two test locations are more common than the other two locations 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()
```



```
Neonics$Endpoint <- as.factor(Neonics$Endpoint)
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: The two most common endpoints are NOEL and LOEL. NOEL is defined as "No-observable-effect-level: highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test." LOEL is defined as "Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls."

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"
```

```r
# collectDate is a character, so I will change it to a date.

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

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

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

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

```r
# Litter was sampled in on August 2 and August 30, 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`?

```r
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"
```

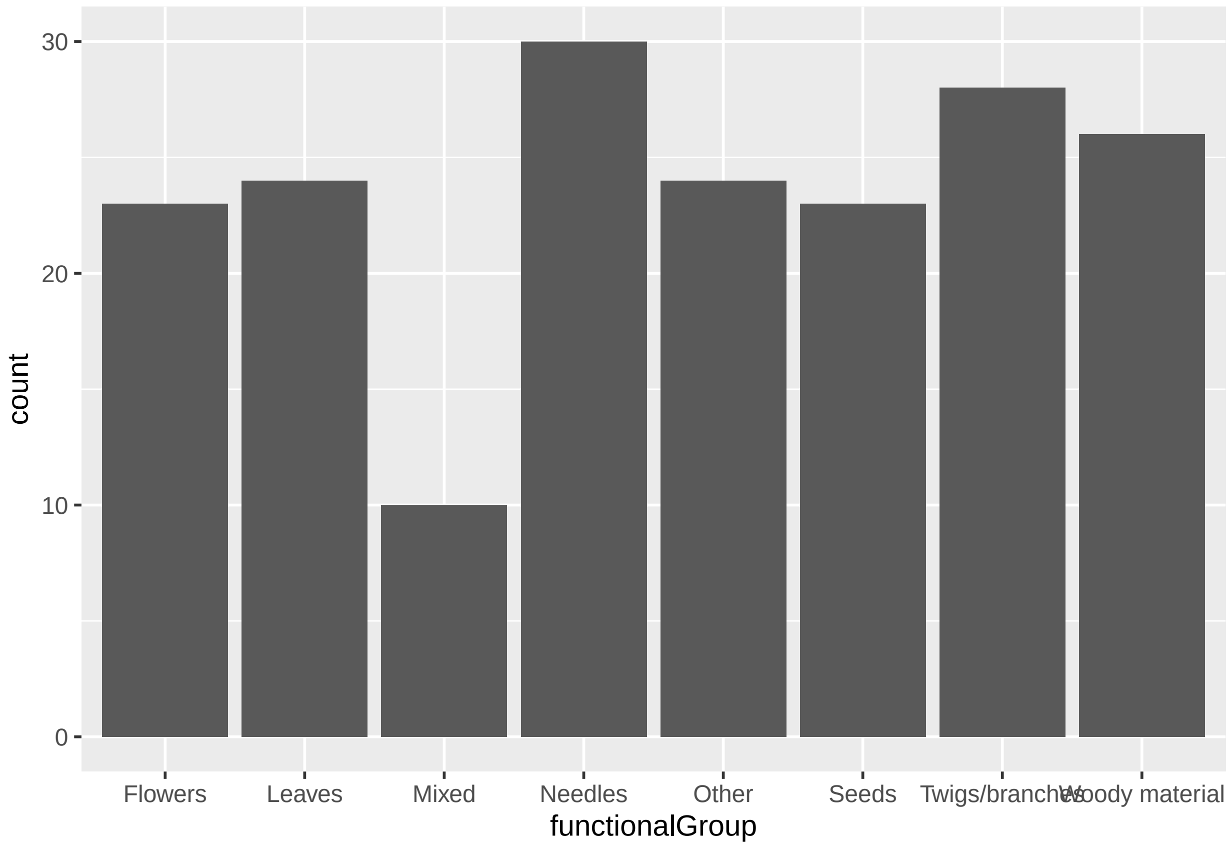
```r
summary(Litter$plotID)
```

```
##    Length     Class      Mode 
##       188 character character
```

> Answer: 12 plots were sampled at Niwot Ridge. The 'unique' function lists all of the unique plot ID values (i.e., the name of each unique plot ID), whereas the 'summary' funtions tells you how many plot ID observations there are (i.e., the length), the class, and the mode.

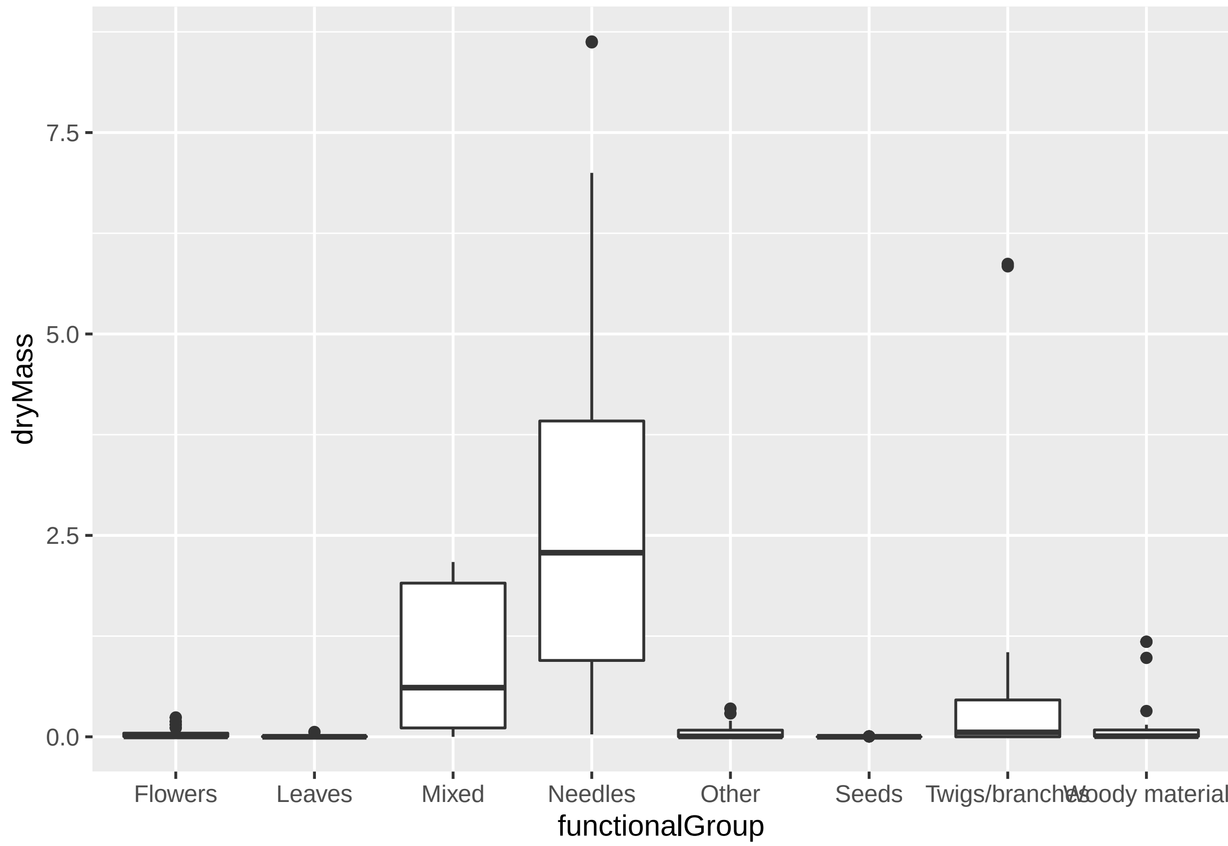
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

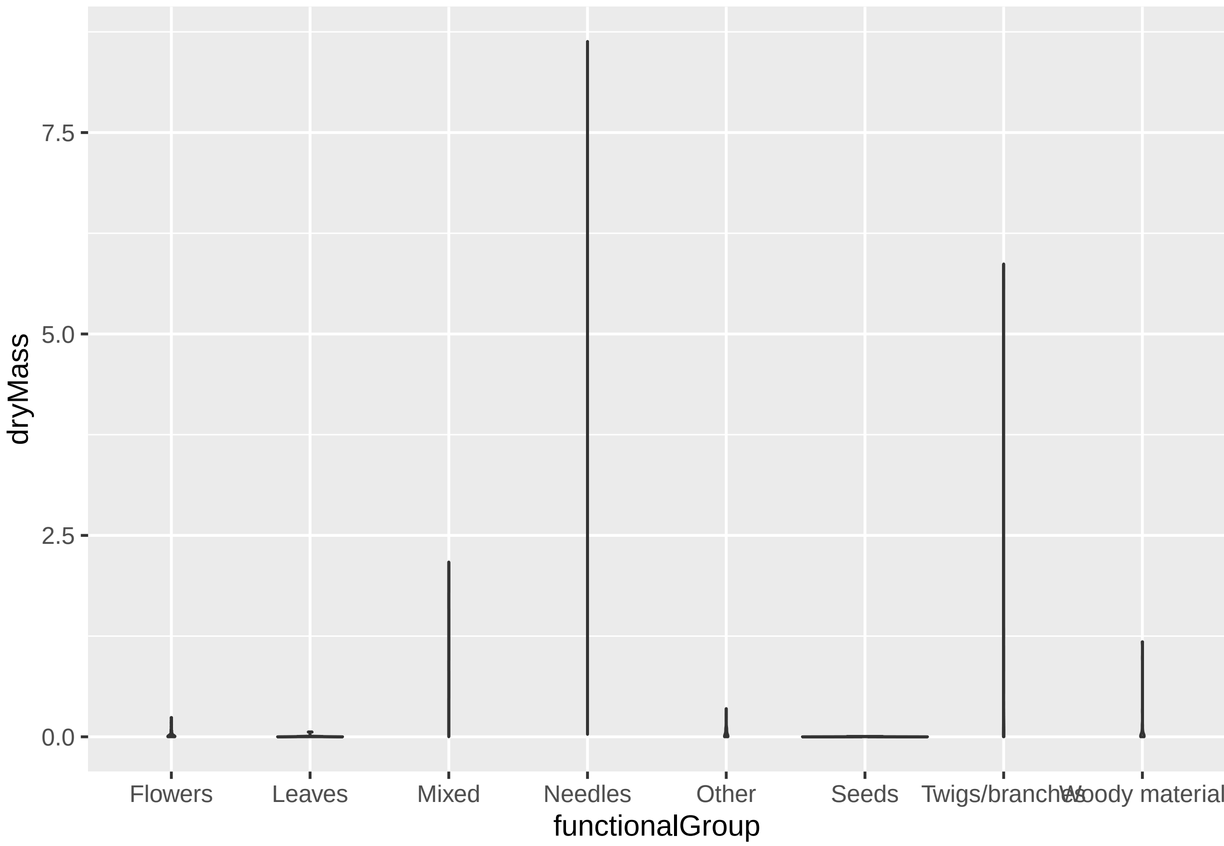


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




```r
# violin plot
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass))
```



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

> Answer: The boxplot is a more effective visualization option than the violin plot in this case because, for many of the functional groups, the 'dryMass' data has a tight distribution/little variation. Therefore, a more complex visualization, such as the violin plot, is not as useful.

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

> Answer: Needles