Assignment 3: Data Exploration

Leyi Ruan

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.

```
library(tidyverse)
getwd()
```

```
## [1] "/Users/ruanleyi/Documents/Year2/872 Environmental Data Analytics/Environmental_Data_Analytics_2
```

```
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: They are absorbed by plants and can be present in pollen and nectar, making them toxic to bees. There was uncertainty about the impact these insecticides were having on 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: Plant debris, including woody debris and litter, is an essential but frequently overlooked component of carbon (C) storage in forest ecosystems.

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: * The litterfall and fine woody debris sampling data products provide mass data for plant functional groups from individual sampling bouts. * Spatial Sampling Design * Temporal Sampling Design

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(as.factor(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 common effect that are studied

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(as.factor(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. They are all in the bees category. They are of interest over other insects because bees are greatly impacted by the intesticides.

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

```
summary(as.factor(Neonics$Conc.1..Author.))
```

```
##    0.37/      10/      NR/       NR        1     1023    0.40/       2/
##      208      127      108       94       82       80       69       63
##       10   0.053/      100      50/     0.5/     0.03    0.05/     0.45
##       62       59       56       51       45       44       43       43
##     0.1/    0.45/     1.0/    2.27/       50    0.125     500/      0.5
##       42       40       40       40       36       33       33       32
##   0.048/    0.15/       1/       48    25.0/      12/    0.027      2.4
##       30       30       30       30       28       27       26       26
##     0.2/    0.56/     100/        3    0.01/    1000/       3/    0.336
##       25       24       23       23       22       22       22       21
##     1.5/     0.05      1.5    2.60/    20.0/        6    6.80/    62.5/
##       21       20       20       20       20       20       20       20
##    0.005     0.4/    0.18/     0.3/     1000       40 0.00355/      0.1
##       18       18       17       17       17       17       16       16
##      0.4     150/      300      80/    0.053     0.24     0.28     125/
##       16       16       16       16       15       15       15       15
##        9   0.0001  0.0004/   0.084/     0.15      0.6    12.5/   144.0/
```

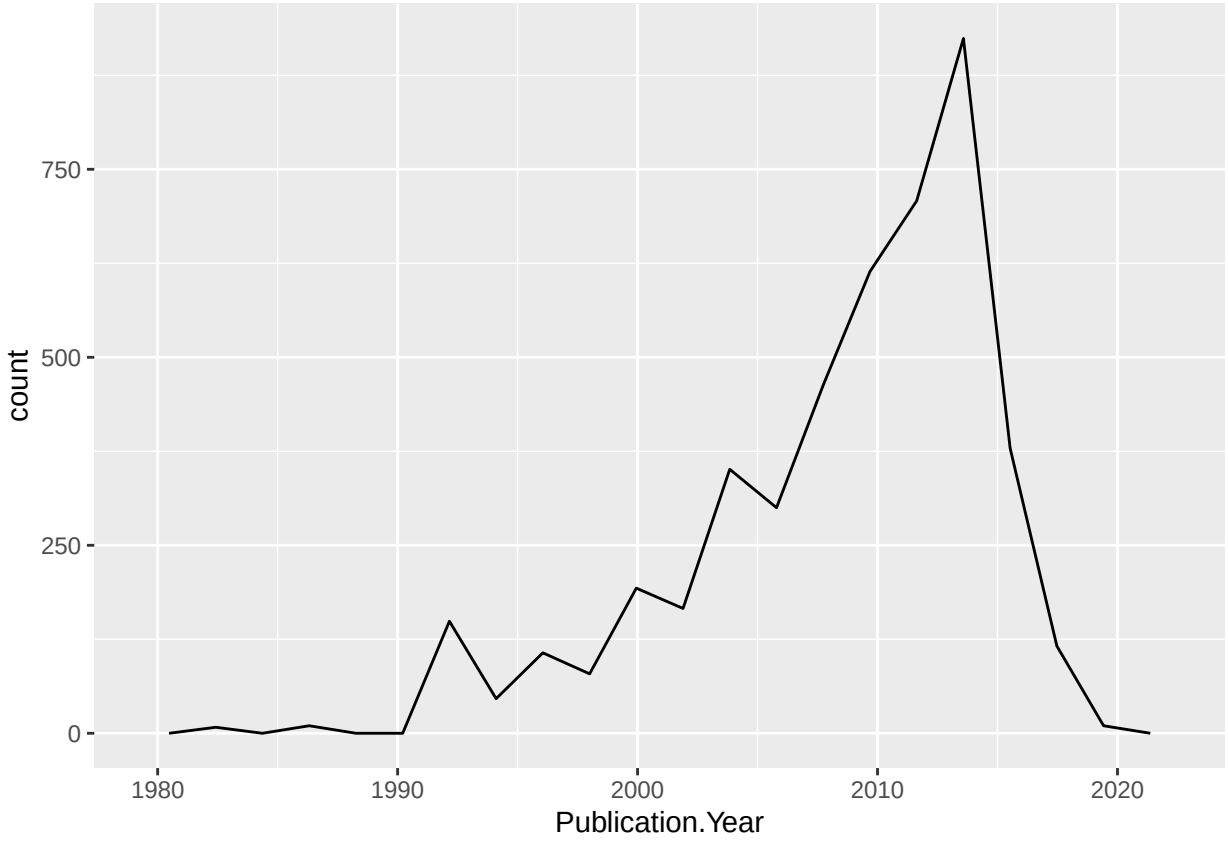
```
##       15       14       14       14       14       14       14       14
##     350/    40.0/       48/       56       84/    0.17/      125       14
##       14       14       14       14       14       13       13       13
##       16       17   0.047/    0.25/    0.28/    1.28/    1.81/      112
##       13       13       12       12       12       12       12       12
##      150      2.5/       25       60/       75/    0.02/   0.025/     0.29
##       12       12       12       12       12       11       11       11
##    37.5/       4/        5  (Other)
##       11       11       11     1817
```

Answer: Conc.1..Author is character variable. It is not numeric because concentrains are based on various levels.

Explore your data graphically (Neonics)

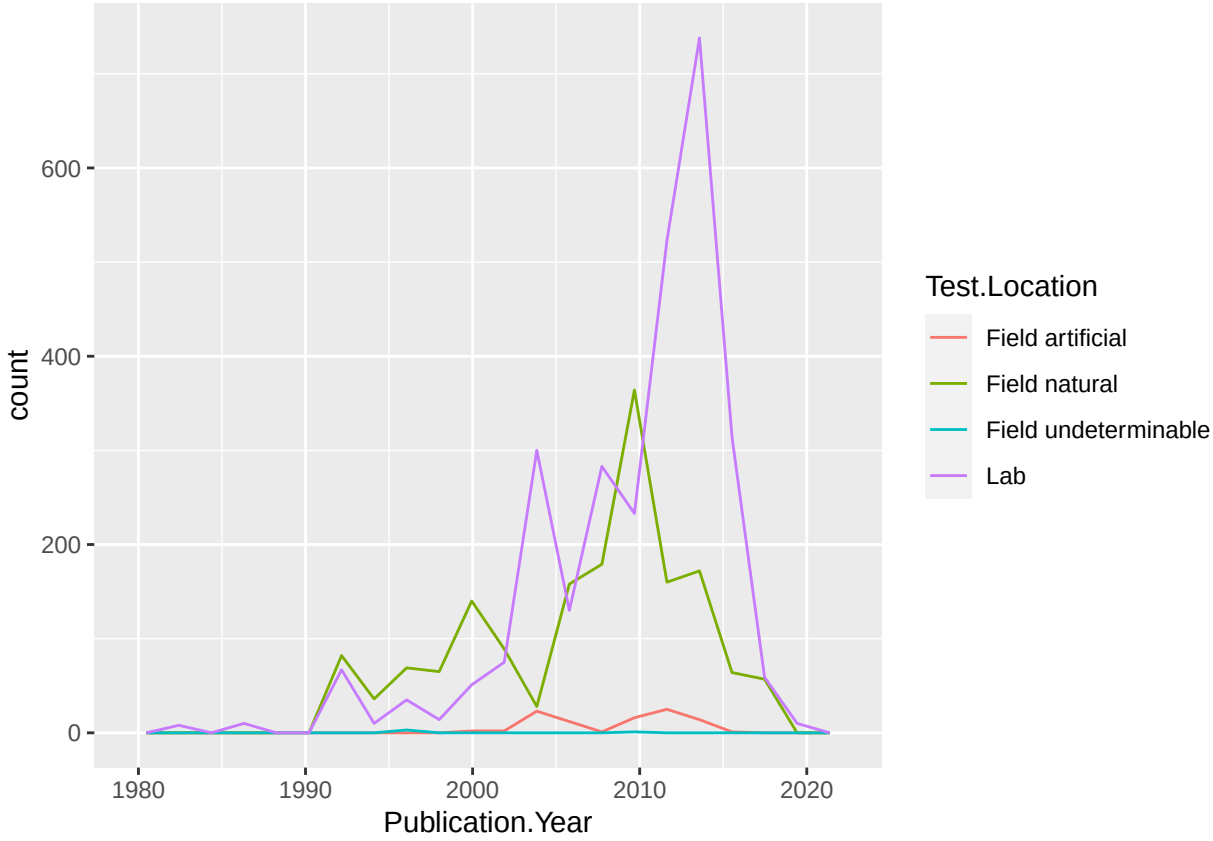
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics)+
  geom_freqpoly(aes(x=Publication.Year), bins= 20)
```



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= 20)
```

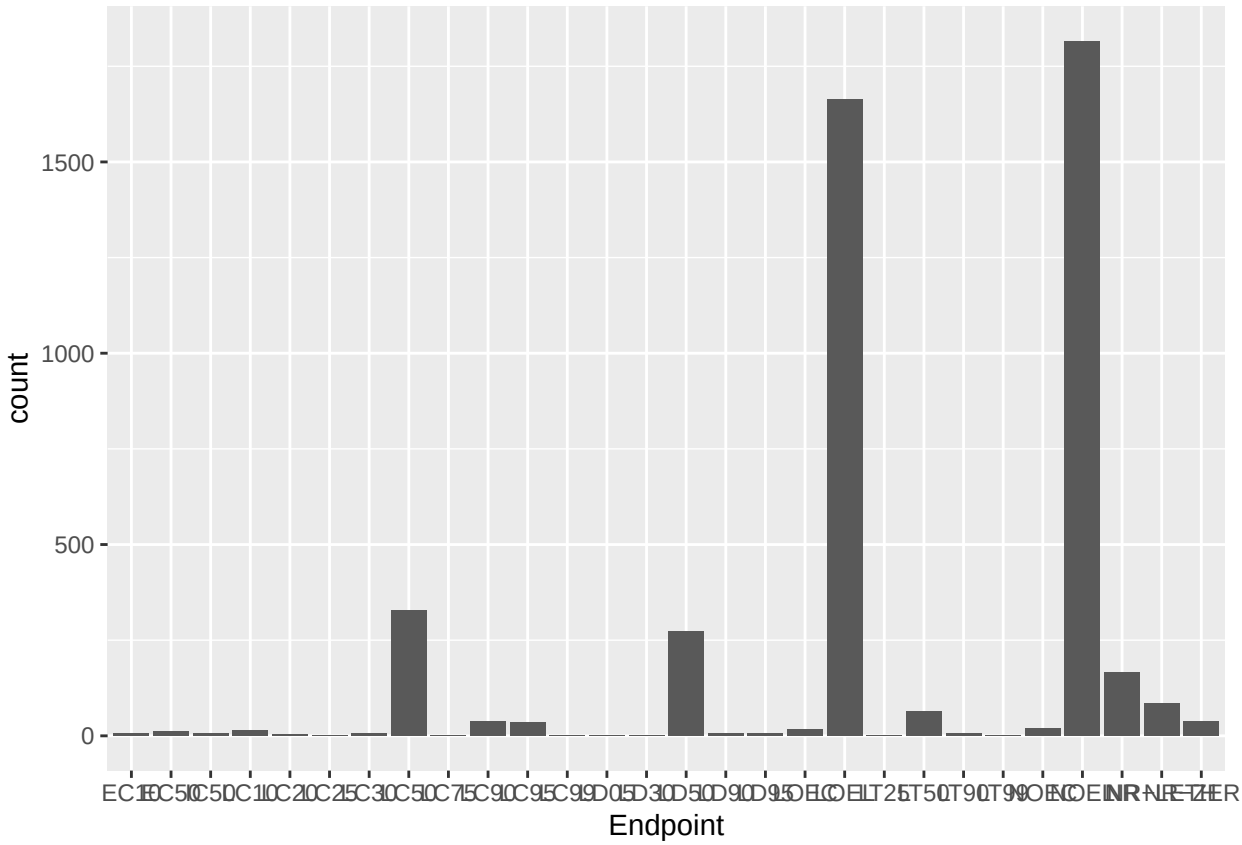


Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: Lab is the most common test locations. They do differ over time, difference between number of tests conducted in Lab and other locations are getting larger.

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()
```




```
summary(as.factor(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: LOEL and NOEL are two most common points. They are defined lowest-observable-effect-level, and no-observed-effects-residue.

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"
```

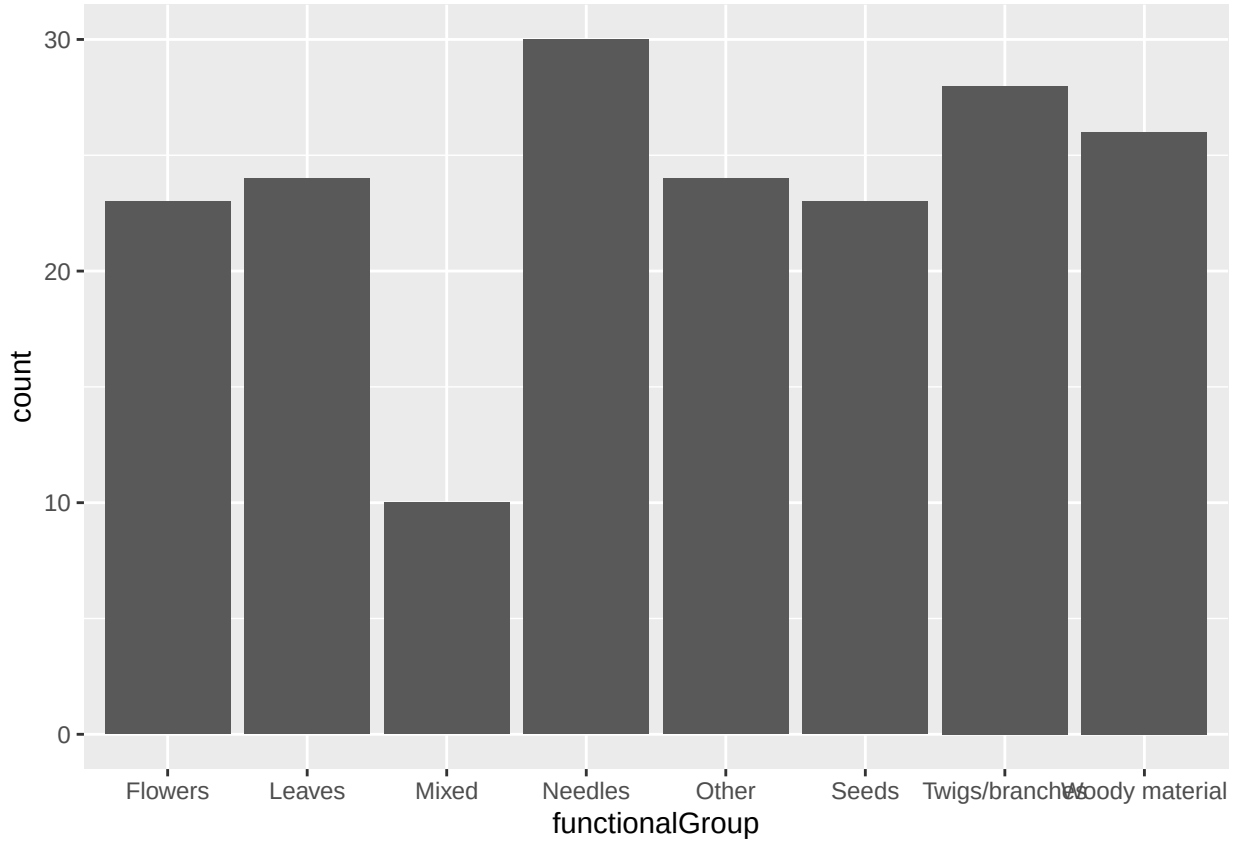
```
summary(Litter$plotID)
```

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

Answer: unique function gives the values that plotID can be; summary function returns the class, mode, and length of plotID, which is not what we need.

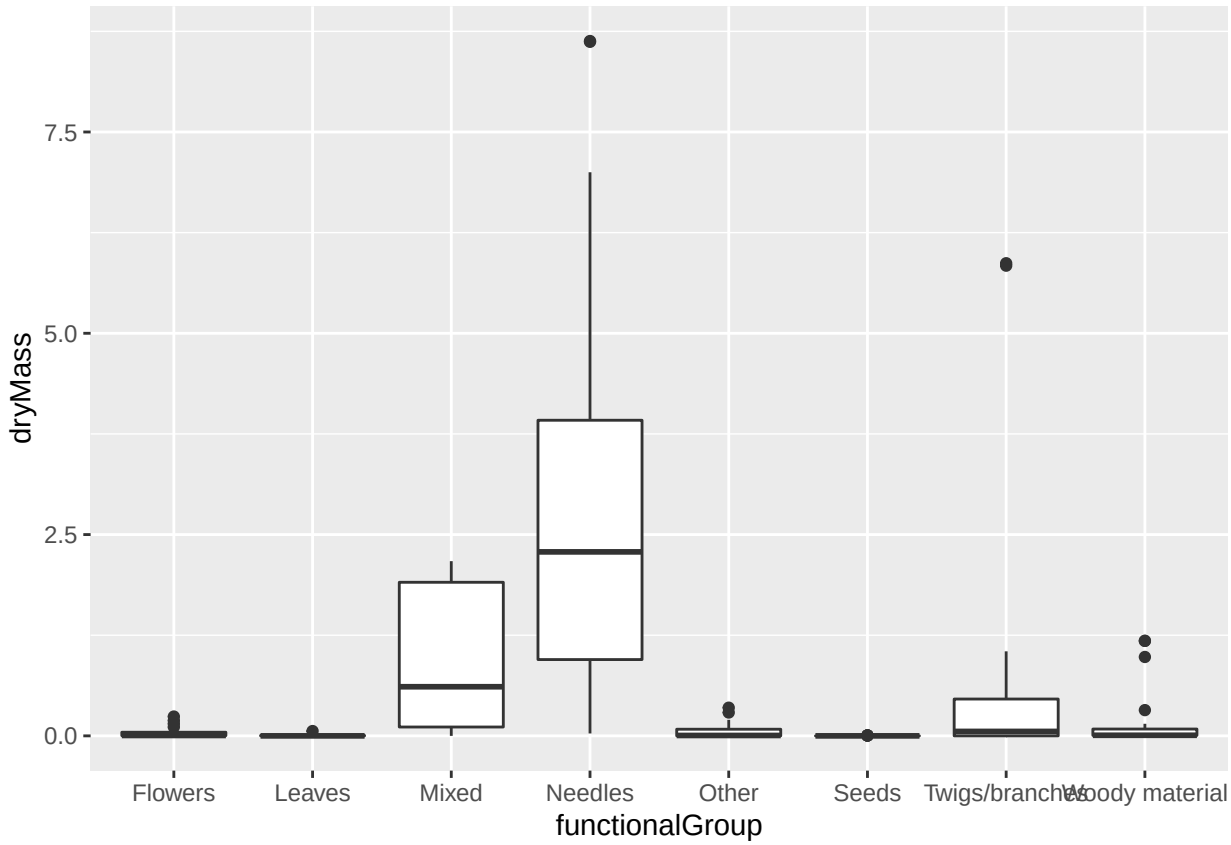
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
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.

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

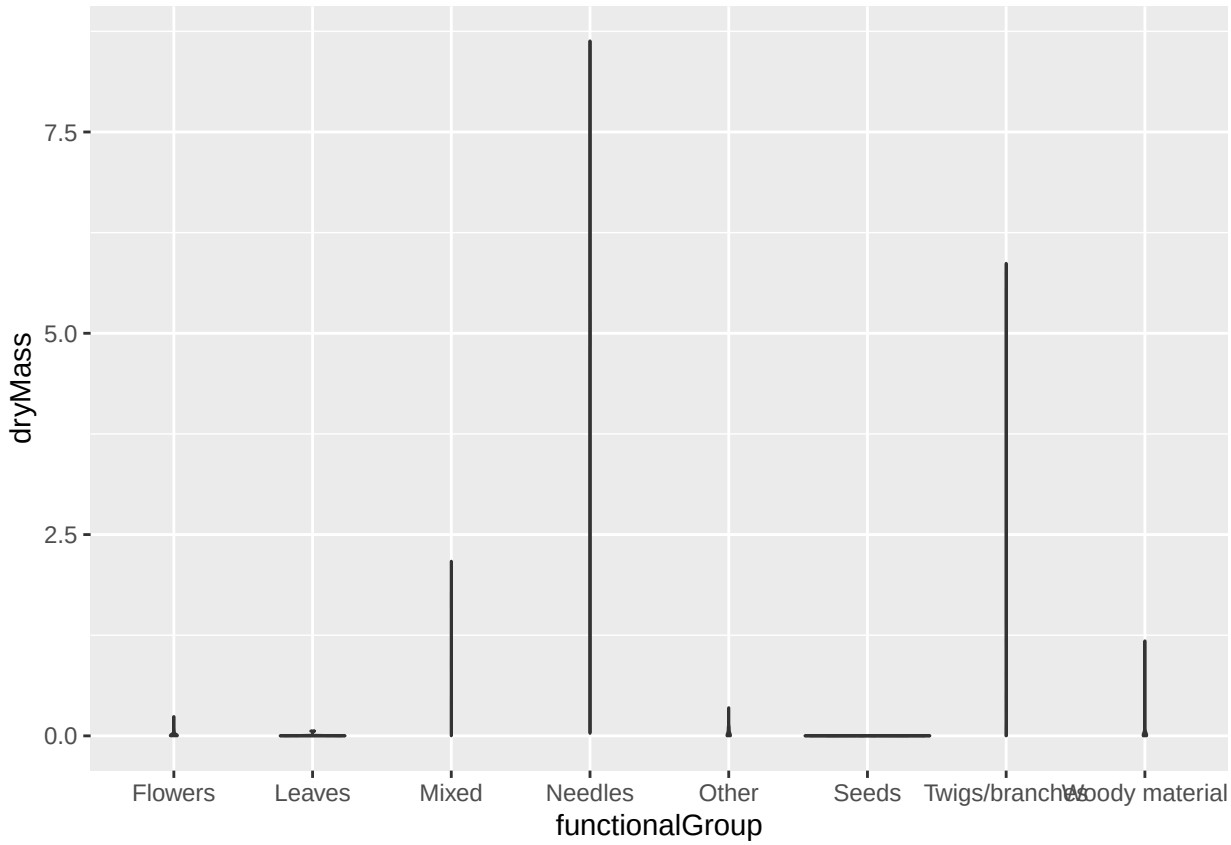



```
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75))
```

```
## 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: because violin chart cannot reflect how the drymass data is distributed on each functional group.

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

Answer: Needles