Assignment 3: Data Exploration

Abby Liu

Spring 2023

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. Assign a useful **name to each code chunk** and include ample **comments** with your code.
5. Be sure to **answer the questions** in this assignment document.
6. When you have completed the assignment, **Knit** the text and code into a single PDF file.
7. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai.

TIP: If your code extends past the page when knit, tidy your code by manually inserting line breaks.

TIP: If your code fails to knit, check that no `install.packages()` or `View()` commands exist in your code.

Set up your R session

1. Check your working directory, load necessary packages (tidyverse, lubridate), 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.

```
# 1. Set up working directory
getwd()
```

```
## [1] "/home/guest/EDA-Spring2023"
```

```
# 2. Load packages
library(tidyverse)
library(lubridate)

# 3. Import datasets
ecotox <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = T)
litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv", stringsAsFactors = T)
```

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: According to scientific research, neonicotinoids can pose undocumented threats to food webs. In particular, researchers have found that neonicotinoids can move from treated plants to pollinators and from plants to pests to natural enemies. Worse, transmission through simple food chains portends widespread, undocumented transmission into entire food webs. Although further research is needed to document the ecosystem-wide transmission and consequences of neonicotinoids to establish their true costs and benefits, serious efforts must be made to decrease the scale of their use. Source: https://www.pnas.org/doi/10.1073/pnas.2017221117

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: Woody debris is an important part of forest and stream ecosystems because it has a role in carbon budgets and nutrient cycling, is a source of energy for aquatic ecosystems, provides habitat for terrestrial and aquatic organisms, and influences water flows and sediment transport. Also, fine woody debris may act as a tinder that promotes the start and spreading of forest fires. Sampling the extent of fine woody debris indicates fire risk and gives clues about the rate of forest decay. Sources: https://www.fs.usda.gov/research/treesearch/20001; https://placebasedbasics.weebly.com/fine-and-course-woody-debris-analysis.html#:~:text=Fine%20woody%20debris%20may%20act,the%20rate%20of%20forest%20decay.

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. All masses are reported at the spatial resolution of a single trap and the temporal resolution of a single collection event. 2. The sampling plots differ based on vegetation density in tower plots and there are a set of rules that restrict the number of plots can be sampled. 3. Trap placement within plots may be either targeted or randomized, depending on the vegetation.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
# Use dim() to get the dimension of each dataset
dim(ecotox)
```

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

```
dim(litter)
```

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

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?

```
# Use summary() to get basic information of the dataset and use sort() to sort
# them in descending order
sort(summary(ecotox$Effect), TRUE)
```

```
##       Population        Mortality         Behavior Feeding behavior
##             1803             1493              360              255
##     Reproduction      Development        Avoidance         Genetics
##              197              136              102               82
##        Enzyme(s)           Growth       Morphology    Immunological
##               62               38               22               16
##     Accumulation     Intoxication     Biochemistry          Cell(s)
##               12               12               11                9
##       Physiology        Histology       Hormone(s)
##                7                5                1
```

Answer: The most common effects that are studied are population, mortality, behavior, feeding behavior and reproduction. Mortality measures the death cause by direct impact of chemicals and population measures the change in numbers of certain species, which altogether indicate the magnitude of the impact of use of neonicotinoids. The behavior related effects measure the affect that neonicotinoids can have on species.

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.[TIP: The `sort()` command can sort the output of the summary command...]

```
# Use summary() to get basic information of the dataset and use sort() to sort
# them in descending order
sort(summary(ecotox$Species.Common.Name), TRUE)
```

```
##                          (Other)                       Honey Bee
##                              670                             667
##                    Parasitic Wasp            Buff Tailed Bumblebee
##                              285                             183
##                Carniolan Honey Bee                      Bumble Bee
##                              152                             140
##                   Italian Honeybee                  Japanese Beetle
##                              113                              94
##                 Asian Lady Beetle                   Euonymus Scale
##                               76                              75
##                         Wireworm                European Dark Bee
##                               69                              66
##                 Minute Pirate Bug              Asian Citrus Psyllid
##                               62                              60
```

```
##                   Parastic Wasp            Colorado Potato Beetle
##                              58                                57
##                 Parasitoid Wasp              Erythrina Gall Wasp
##                              51                                49
##                    Beetle Order       Snout Beetle Family, Weevil
##                              47                                47
##          Sevenspotted Lady Beetle                 True Bug Order
##                              46                                45
##             Buff-tailed Bumblebee                   Aphid Family
##                              39                                38
##                  Cabbage Looper            Sweetpotato Whitefly
##                              38                                37
##                  Braconid Wasp                  Cotton Aphid
##                              33                                33
##                 Predatory Mite            Ladybird Beetle Family
##                              33                                30
##                     Parasitoid                   Scarab Beetle
##                              30                                29
##                  Spring Tiphia                   Thrip Order
##                              29                                29
##             Ground Beetle Family              Rove Beetle Family
##                              27                                27
##                  Tobacco Aphid                  Chalcid Wasp
##                              27                                25
##           Convergent Lady Beetle                 Stingless Bee
##                              25                                25
##                Spider/Mite Class              Tobacco Flea Beetle
##                              24                                24
##                Citrus Leafminer                  Ladybird Beetle
##                              23                                23
##                     Mason Bee                         Mosquito
##                              22                                22
##                  Argentine Ant                           Beetle
##                              21                                21
##        Flatheaded Appletree Borer             Horned Oak Gall Wasp
##                              20                                20
##               Leaf Beetle Family                Potato Leafhopper
##                              20                                20
##        Tooth-necked Fungus Beetle                  Codling Moth
##                              20                                19
##         Black-spotted Lady Beetle                  Calico Scale
##                              18                                18
##              Fairyfly Parasitoid                   Lady Beetle
##                              18                                18
##           Minute Parasitic Wasps                     Mirid Bug
##                              18                                18
##                Mulberry Pyralid                      Silkworm
##                              18                                18
##                  Vedalia Beetle             Araneoid Spider Order
##                              18                                17
##                     Bee Order                  Egg Parasitoid
##                              17                                17
##                   Insect Class          Moth And Butterfly Order
##                              17                                17
```

```
##       Oystershell Scale Parasitoid Hemlock Woolly Adelgid Lady Beetle
##                                 17                                16
##              Hemlock Wooly Adelgid                              Mite
##                                 16                                16
##                         Onion Thrip              Western Flower Thrips
##                                 16                                15
##                        Corn Earworm                  Green Peach Aphid
##                                 14                                14
##                           House Fly                         Ox Beetle
##                                 14                                14
##                  Red Scale Parasite                 Spined Soldier Bug
##                                 14                                14
##              Armoured Scale Family                   Diamondback Moth
##                                 13                                13
##                       Eulophid Wasp                  Monarch Butterfly
##                                 13                                13
##                       Predatory Bug              Yellow Fever Mosquito
##                                 13                                13
##                 Braconid Parasitoid                       Common Thrip
##                                 12                                12
##       Eastern Subterranean Termite                             Jassid
##                                 12                                12
##                          Mite Order                          Pea Aphid
##                                 12                                12
##                   Pond Wolf Spider           Spotless Ladybird Beetle
##                                 12                                11
##             Glasshouse Potato Wasp                          Lacewing
##                                 10                                10
##             Southern House Mosquito            Two Spotted Lady Beetle
##                                 10                                10
##                          Ant Family                       Apple Maggot
##                                  9                                 9
```

Answer: The six most commonly studied species are honey bee, parasitic wasp, buff tailed bumblebee, carniolan honey bee, bumble bee, Italian honeybee and other unspecified species. The are all in the bee families, which are pollinators. They play important roles in the food webs as they can get the chemicals from treated plants and passed on to other plants through pollination.

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

```
# Use class() to check class
class(ecotox$Conc.1..Author.)
```

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

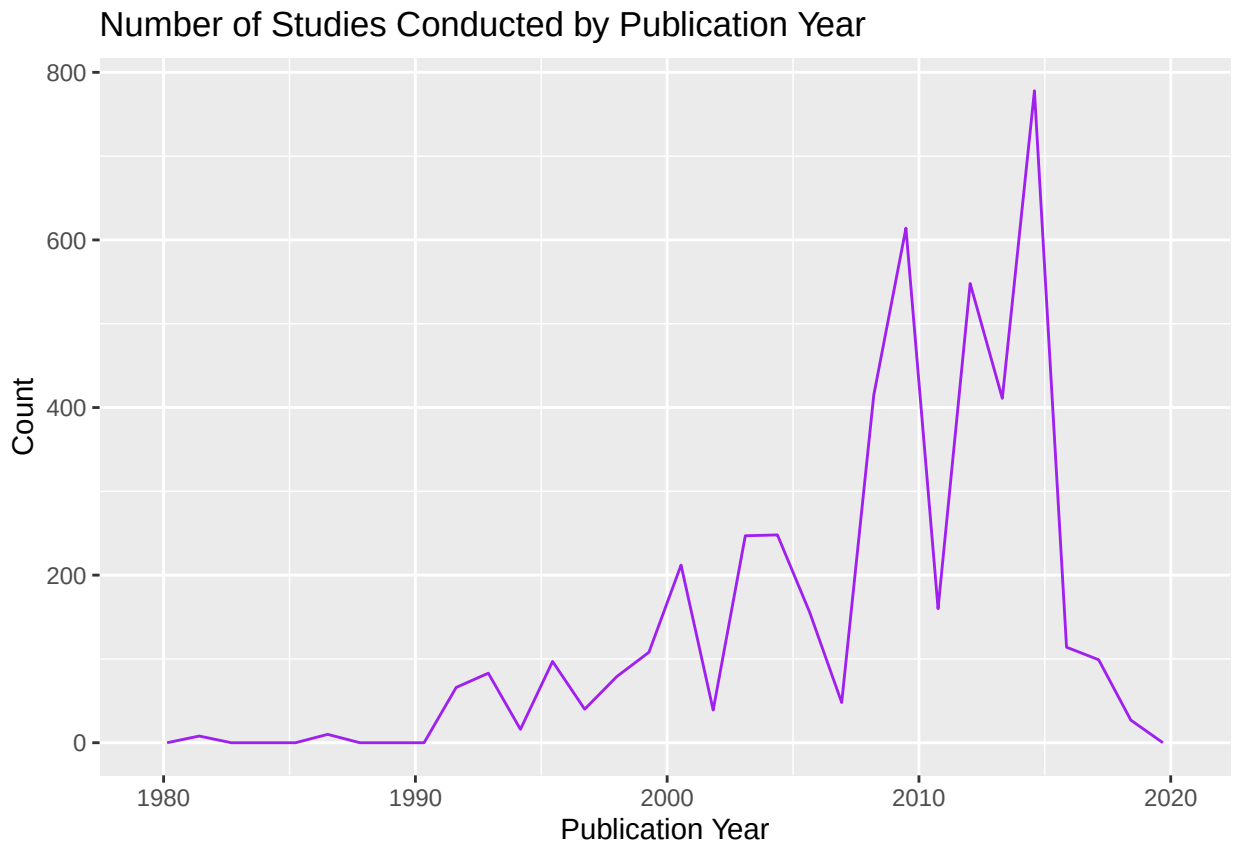
Answer: The class of concentrations is factor. Factor in R is a variable used to categorize and store the data, having the ability to store both string and integer data values as levels. It is not numeric because some of the concentration data are not exact numbers e.g. <4, >100.

Explore your data graphically (Neonics)

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

```
# Use geom_freqpoly() to generate a plot of the number of studies conducted by
# publication year
ggplot(ecotox) + geom_freqpoly(aes(x = Publication.Year), color = "purple") + labs(x = "Publication Yea
    y = "Count", title = "Number of Studies Conducted by Publication Year")
```

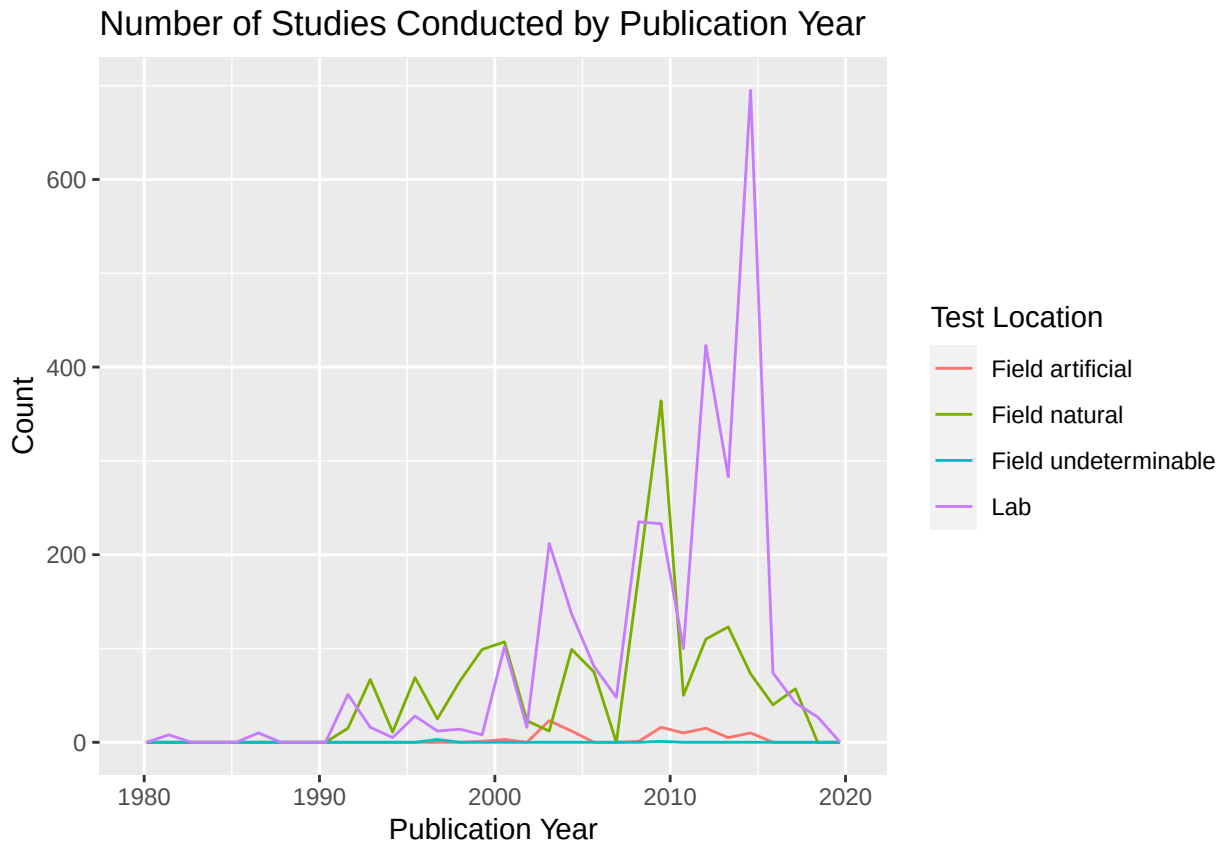
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
# Use geom_freqpoly() to generate a plot of the number of studies conducted by
# publication year and colored by test locations
ggplot(ecotox) + geom_freqpoly(aes(x = Publication.Year, color = Test.Location)) +
    labs(x = "Publication Year", y = "Count", title = "Number of Studies Conducted by Publication Year"
    guides(color = guide_legend(title = "Test Location"))
```

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



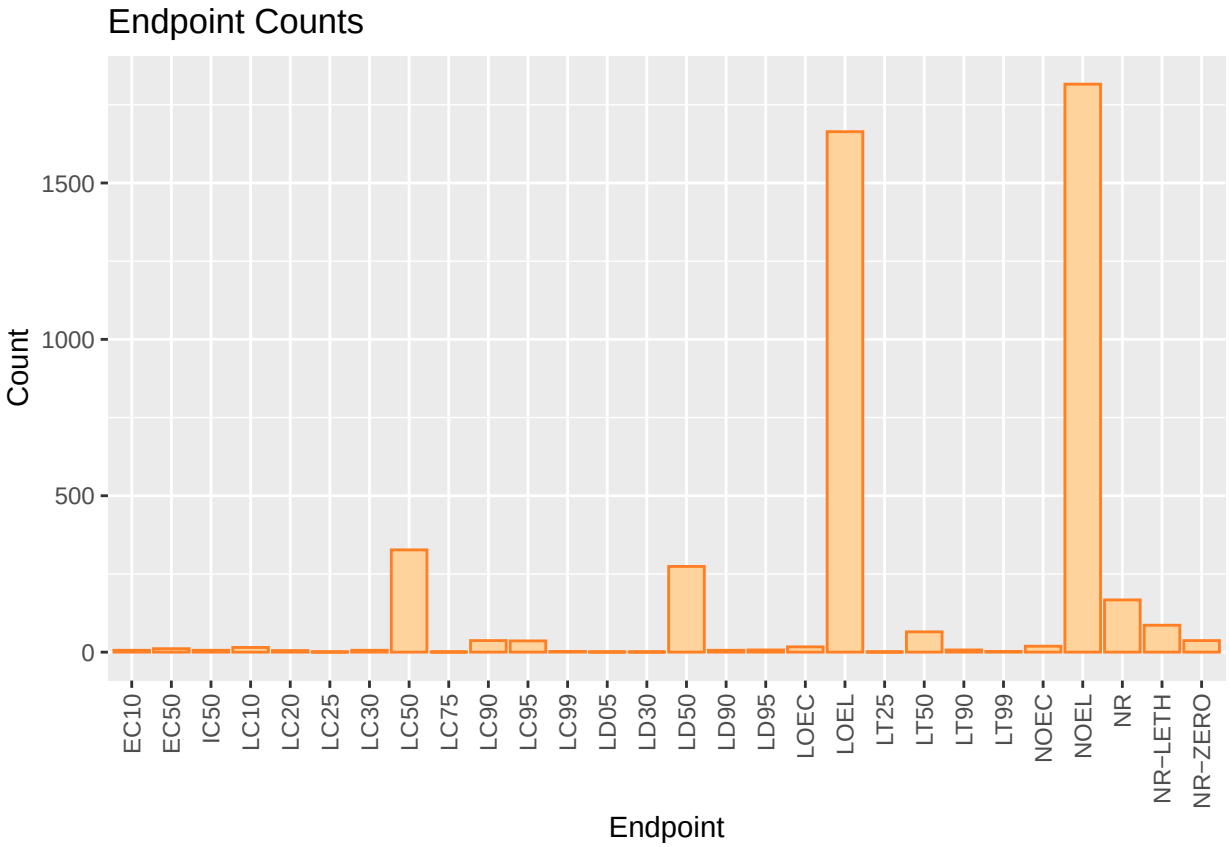
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are lab and field natural. During periods of 1992~2000 and 2008~2010, field natural is the most common test location. Other than those windows, lab remains the most common test location.

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.

[**TIP**: Add `theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))` to the end of your plot command to rotate and align the X-axis labels...]

```
ggplot(ecotox) + geom_bar(aes(x = Endpoint), color = "chocolate1", fill = "burlywood1") +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) + labs(y = "Count",
    title = "Endpoint Counts")
```



Answer: The most common end points are NOEL and LOEL. NOEL (No-observable-effect-level): highest dose producing effects not significantly different from responses of controls according to author's reported statistical test. LOEL (Lowest-observable-effect-level): lowest dose producing effects that were significantly different 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.

```
# 1. Determine the class
class(litter$collectDate)
```

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

```
# 2. Change to date format and confirm the new class
litter$collectDate <- as.Date(litter$collectDate, format = "%Y-%m-%d")
class(litter$collectDate)
```

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

```
# 3. Use unique function to determine the dates that litter was sampled
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`?

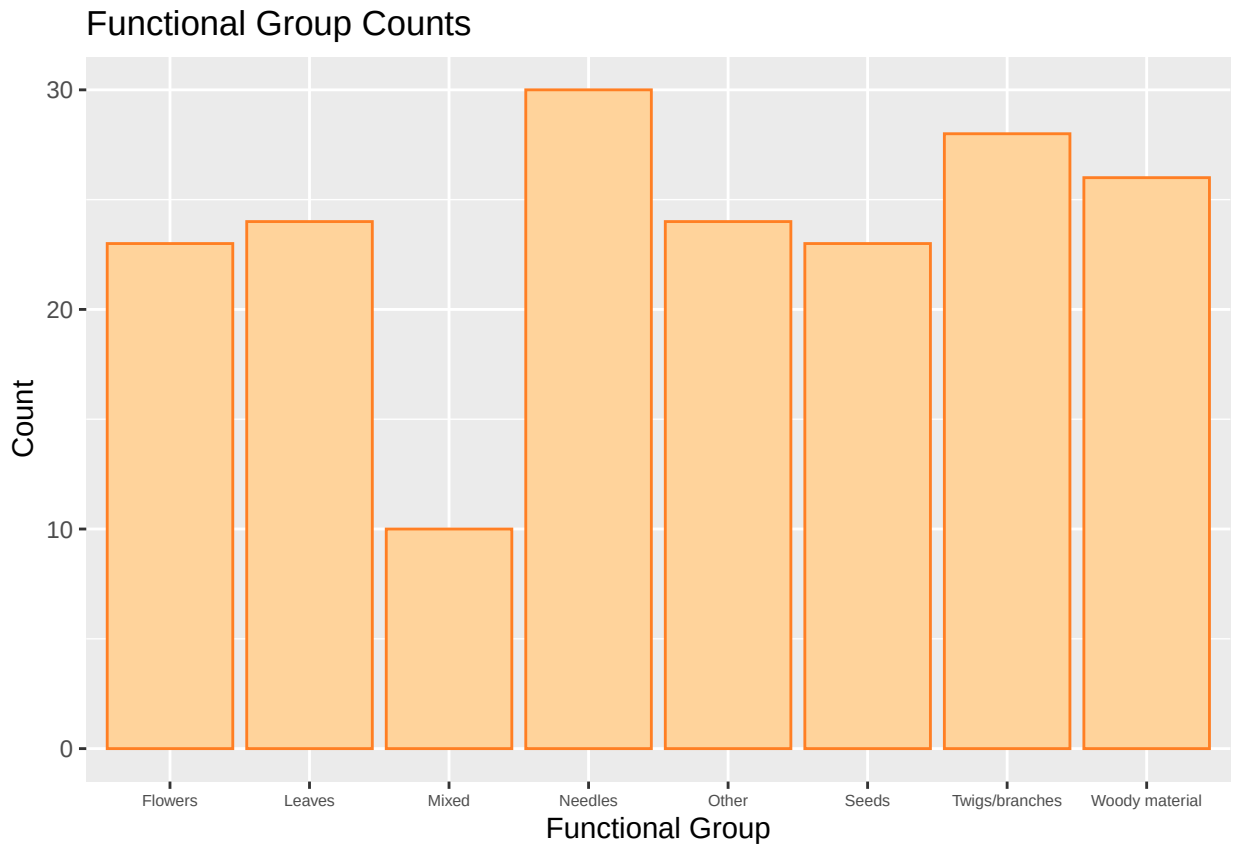
```
unique(litter$plotID)
```

```
##  [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 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: 12 plots were sampled. The information obtained from unique() is solely number of unique values, in this case, plotID. The summary() function provides more detailed information

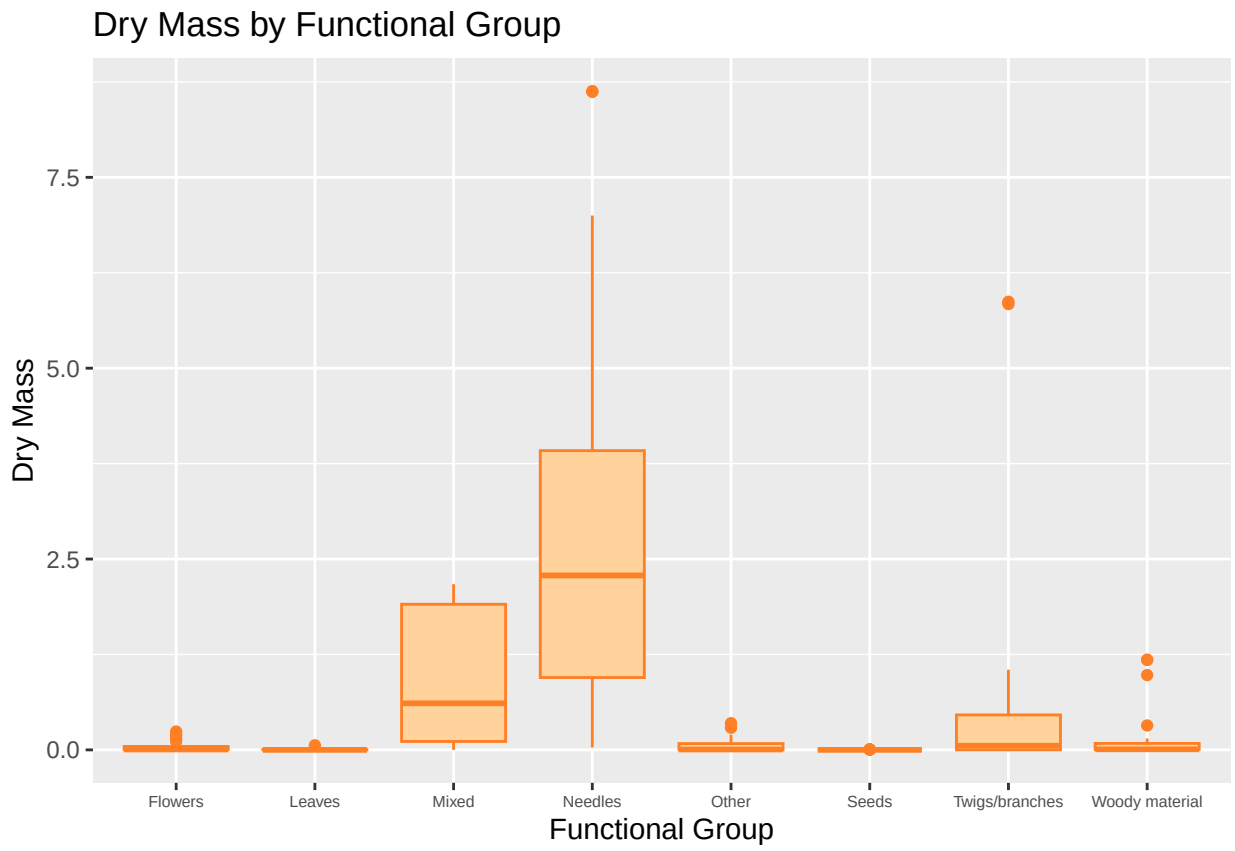
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(litter) + geom_bar(aes(x = functionalGroup), color = "chocolate1", fill = "burlywood1") +
    labs(x = "Functional Group", y = "Count", title = "Functional Group Counts") +
    theme(axis.text.x = element_text(size = 6))
```

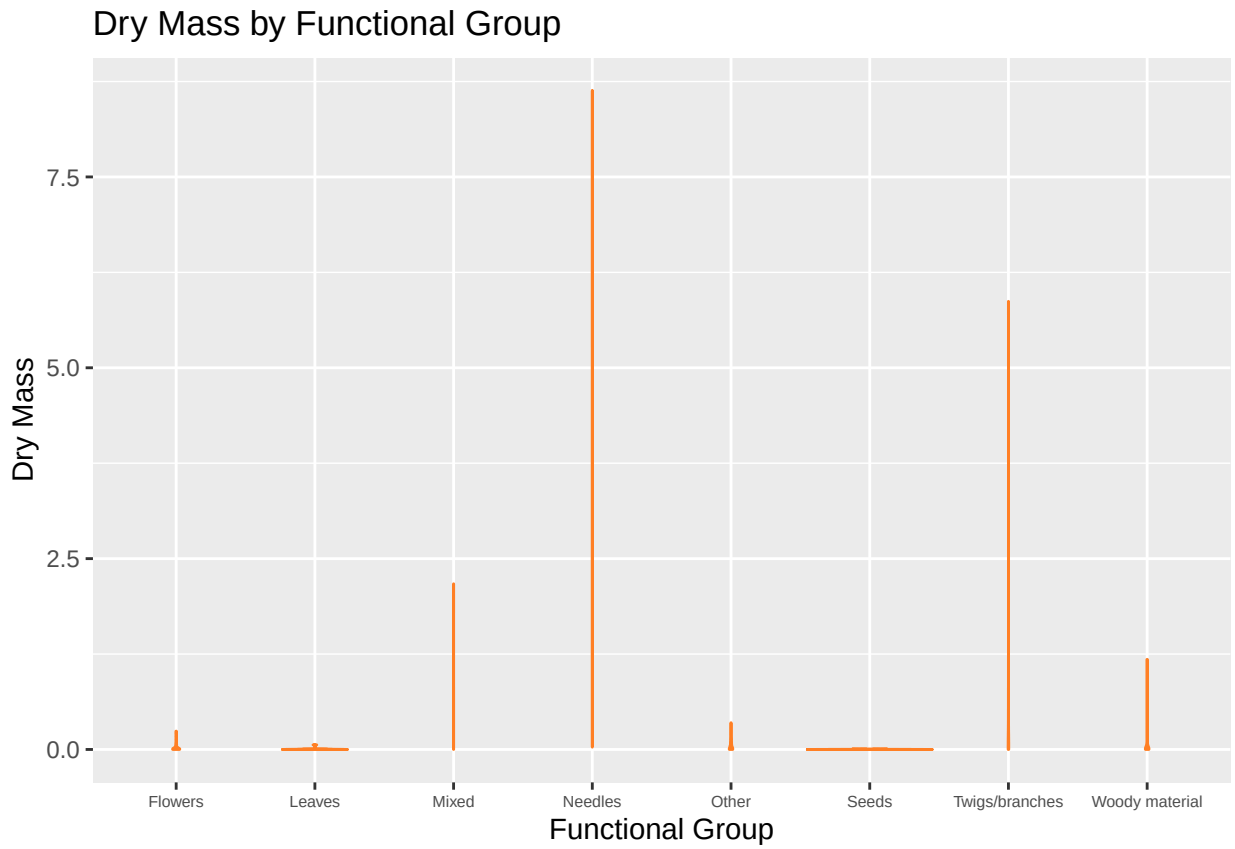


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

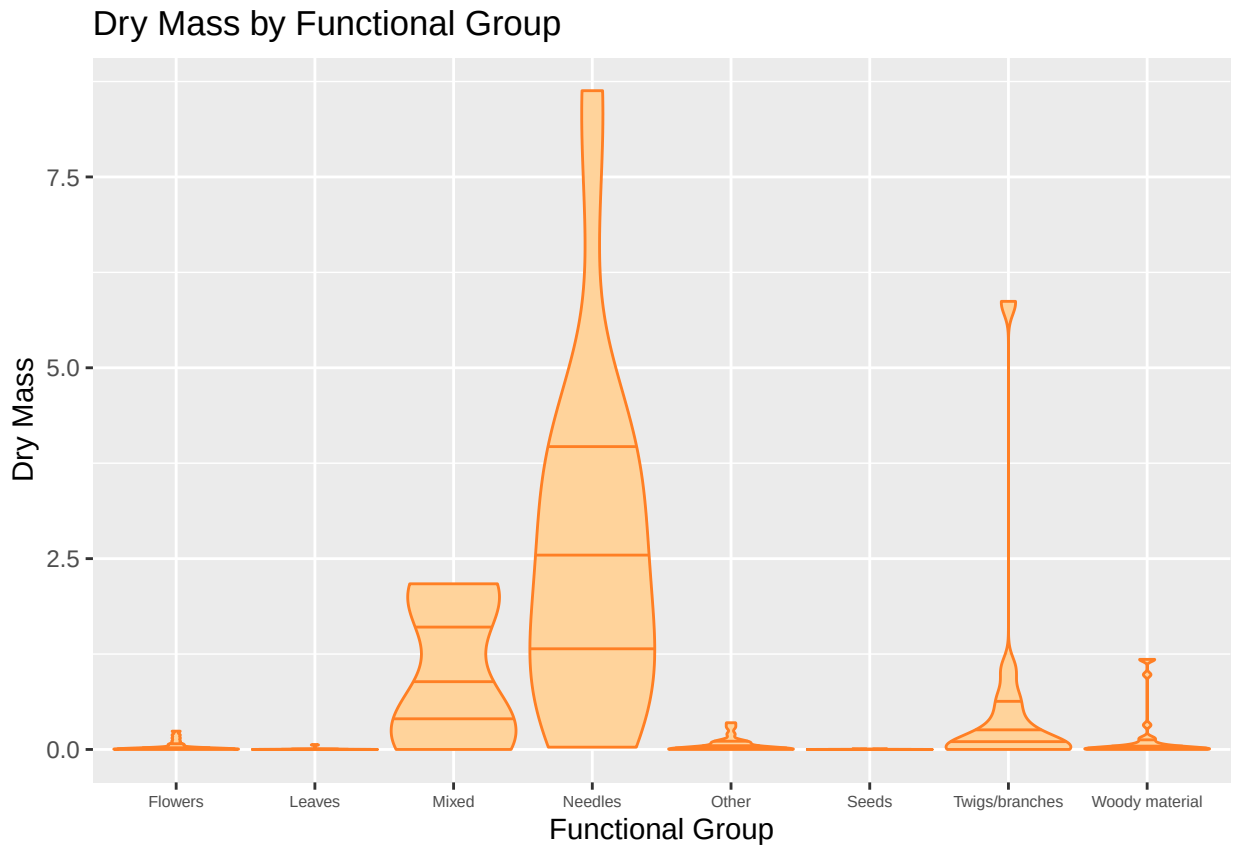
```
# 1. Draw boxplot
ggplot(litter) + geom_boxplot(aes(x = functionalGroup, y = dryMass), color = "chocolate1",
    fill = "burlywood1") + labs(x = "Functional Group", y = "Dry Mass", title = "Dry Mass by Functional
    theme(axis.text.x = element_text(size = 6))
```




```
# 2. Draw violin plot
ggplot(litter) + geom_violin(aes(x = functionalGroup, y = dryMass), draw_quantiles = c(0.25,
    0.5, 0.75), color = "chocolate1", fill = "burlywood1") + labs(x = "Functional Group",
    y = "Dry Mass", title = "Dry Mass by Functional Group") + theme(axis.text.x = element_text(size = 6)
```




```
# 3. Adjust violin plot
ggplot(litter) + geom_violin(aes(x = functionalGroup, y = dryMass), scale = "width",
    draw_quantiles = c(0.25, 0.5, 0.75), color = "chocolate1", fill = "burlywood1") +
    labs(x = "Functional Group", y = "Dry Mass", title = "Dry Mass by Functional Group") +
    theme(axis.text.x = element_text(size = 6))
```




```
# The default scale for violin plot is by area. I've changed the scale to
# 'width' for better visualization, which means all violins have the same
# maximum width.
```

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

> Answer: Ideally, a box plot can only show summary statistics, violin plots can depict summary statistics and the density of each variable. However, in this case, since the default scale for violin plot is by area and the dry mass data of needle function group are much more dispersed than other functional groups, leading to violin plot with extremely skinny shape. Thus, boxplot is a more effective visualization option. Or we can simply tweak the violin plot by changing the scale to 'width' for better visualization, which means all violins have the same maximum width.

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

> Answer: Needles and mixed tend the have the highest biomass at these sites.