Assignment 3: Data Exploration

Sophia Bryson, Section #2

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Exploration.

Directions

1. Change "Student Name, Section #" on line 3 (above) with your name and section number.
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., "FirstLast_A03_DataExploration.Rmd") prior to submission.

The completed exercise is due on 31 January 2022.

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 add the `stringsAsFactors = TRUE` parameter to the function when reading in the CSV files.**

```
#check wd
getwd()
```

```
## [1] "Z:/ENV872/Environmental_Data_Analytics_2022/Assignments"
```

```
#load packages
library(tidyverse)

#load 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: Neonictinoids are a class of insecticides that mimic the structure and function of nicotine. Knowing the toxicity of neonictinoids on insects can (1) be used to know the appropirate amount to use in settings where insecticides may be desired (eg. agriculture) and thereby apply an appropriate amount that will be effective without being excessive and (2) from an ecotoxicology perspective, to assess the impacty of neonictinoids in the environment where their insecticidal effects may be undesireable (eg. a stream ecosystem where benthic macroinvertebrate mortality could lead to a decline in fish species of a higher trophic level dependent on the insects for food).

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: Litter and woody debris function as biomass inputs into the forest floor ecosystem, influencing the movement and cycling or carbon and nutrients.The amount and nature of woody debris also impact the habitat available on the forest floor, as well as the occurrence of fires. They also can be indicators/proxies for primary production and total biomass in a forest.

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 are collected for a vareity of functional groups of litter (leaves, needes, twigs, seeds, etc.) * Spatially, litter sampling occurs in 20 40mx40m plots with 1 to 4 paired elevated and ground traps. * Temporally, ground traps are sampled once per year, while elevated traps vary in frequency from biweekly to bi monthly sampling.

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 "Effect" column, determine the most common effects that are studied. Why might these effects specifically be of interest?

```
summary(Neonics$Effect) %>% sort(decreasing = 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 studied are population and mortality. In addition to being easier to observe across a large number of insects, these effects are the most likely to have siginifcant and direct impacts on broader ecological function and health.

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) %>% sort(decreasing = TRUE) %>% head()
```

```
##                (Other)             Honey Bee        Parasitic Wasp
##                    670                   667                   285
## Buff Tailed Bumblebee   Carniolan Honey Bee            Bumble Bee
##                   183                   152                   140
```

Answer: These six species are all pollinators, which make them interesting for 2 reasons: (1) The loss of pollinators would result in significant repercussions for plant communities, with trophic cascades affecting the rest of the ecosystem which they support. (2) Pollinators may be more like to come into contact with neonictinoids applied in agricultural settings than other 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?

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

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

```
#unique(Neonics$Conc.1..Author.) %>% head(20)
```

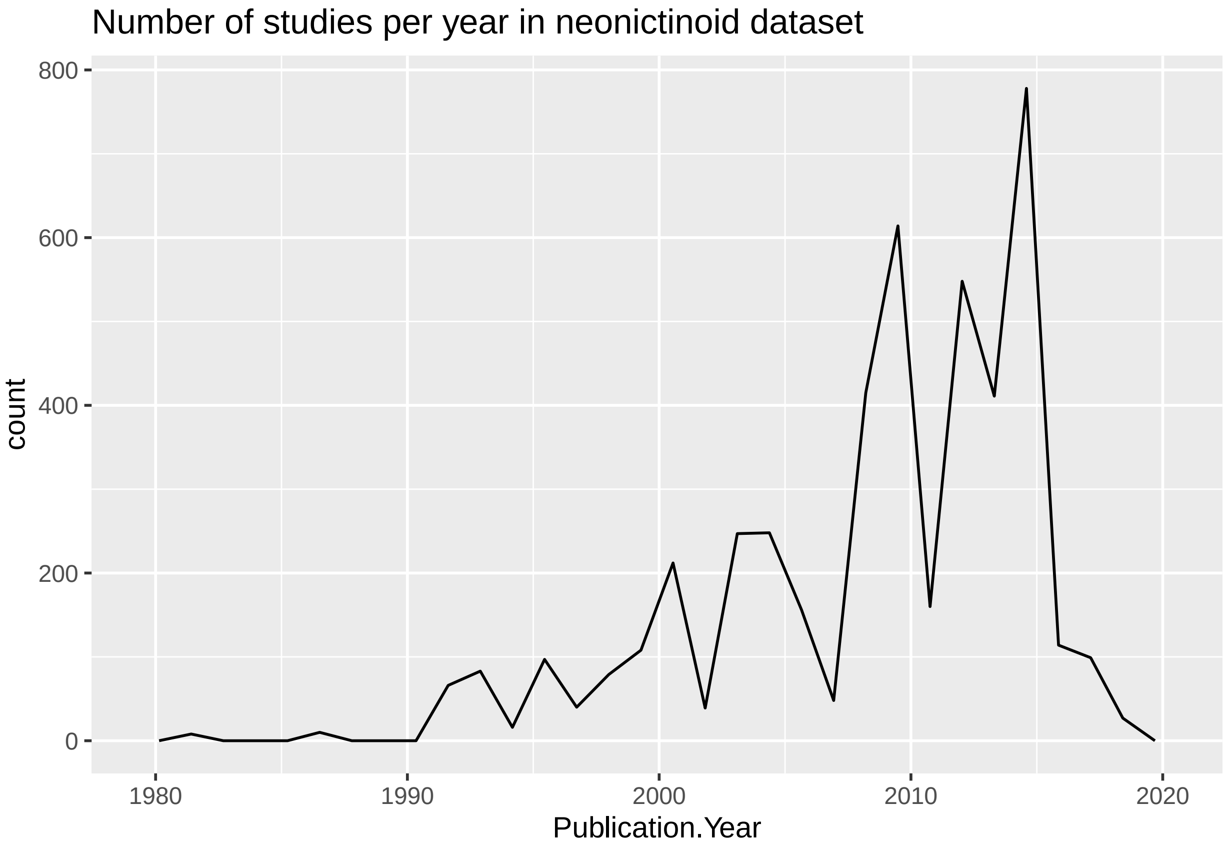
Answer: the "Conc.1..Author" variable is a factor. The proximate reason for this is that we specified that strings should be read as factors when importing the dataset. The ultimate reason for this (ie. the reason this value was a string) is that, at times, the concentration is given as a an approximate value (~10) or as a range/below a given threshhold (<5.00). R coerces columns/vectors to the class that can contain all of the values which they contain, which in this case is a character class.

Explore your data graphically (Neonics)

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

```
neonic_freq <- ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year)) +
                                  labs(title = "Number of studies per year in neonictinoid dataset",
                                       xlab = "Publication year",
                                       ylab = "Number of studies")
neonic_freq
```

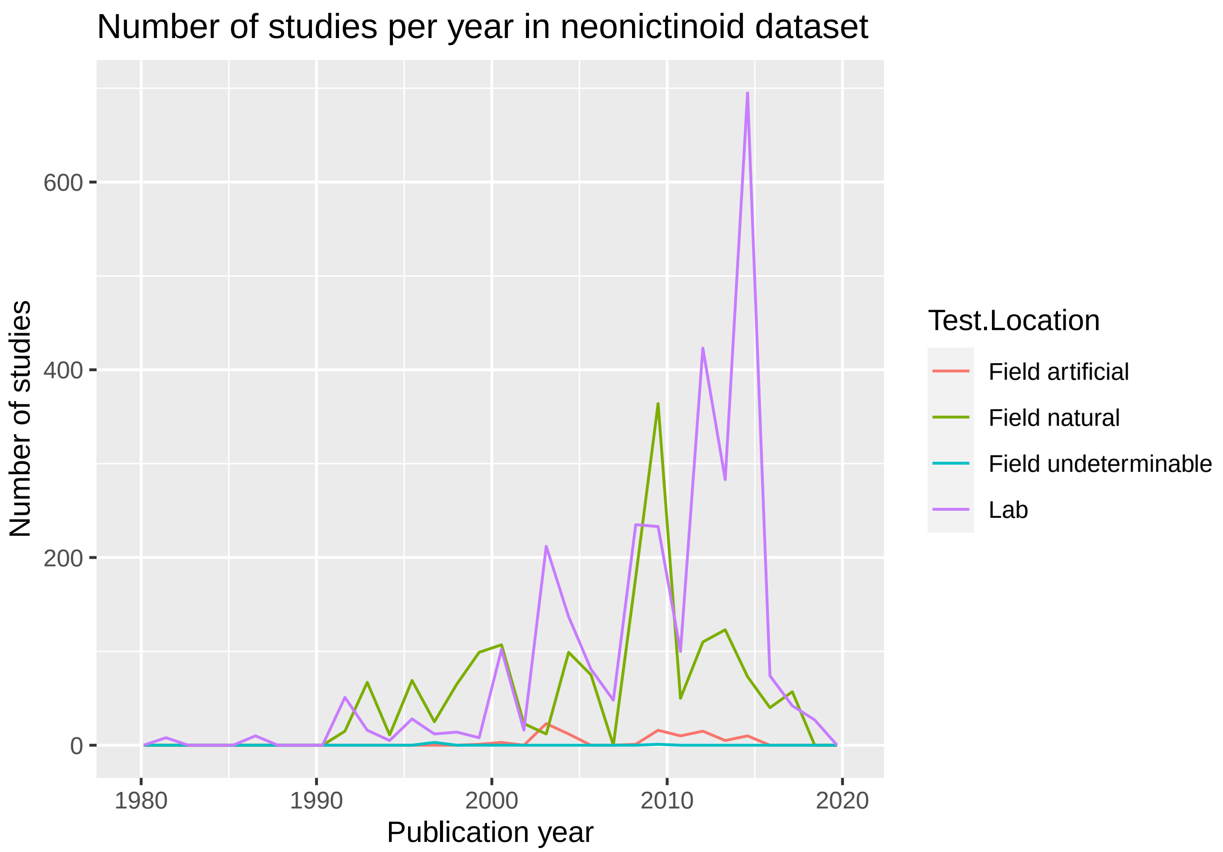
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
neonic_freq <- ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year, color = Test.Location)) +
                                 labs(title = "Number of studies per year in neonictinoid dataset",
                                      x = "Publication year",
                                      y = "Number of studies")
neonic_freq
```

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

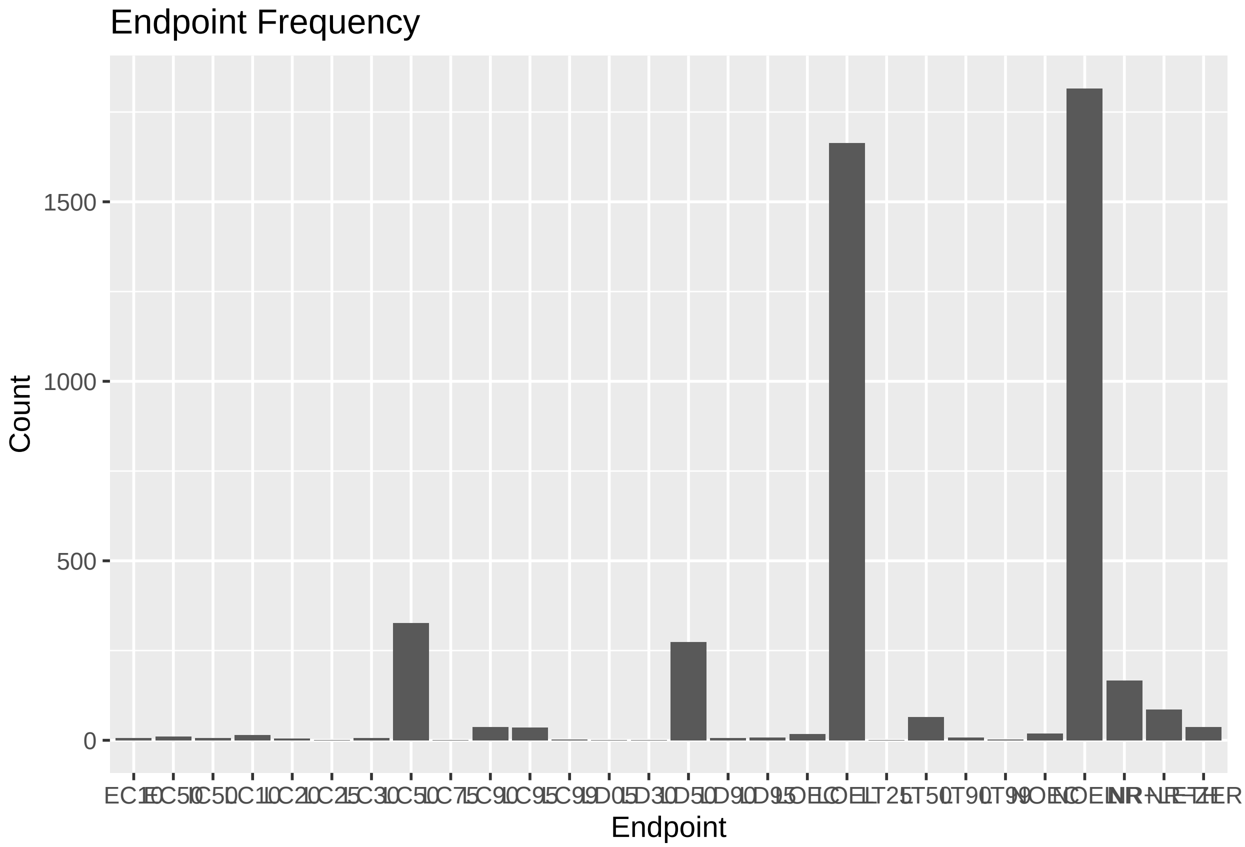


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: Natural field tests and lab tests are the most common, with natural field test predominating between 1990 and 2000 and lab tests growing in frequency particularly post 2010.

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.

```
neonic_endpoints <- ggplot(Neonics) + geom_bar(aes(x = Endpoint)) +
                                      labs(title = "Endpoint Frequency",
                                           x = "Endpoint",
                                           y = "Count")
neonic_endpoints
```



Answer: The two most common endpoints are: 1. NOEL (“no observable effect level”) - this is the highest exposure level with no observable toxic effects 2. LOEL(“lowest observable effect level”) - this is the lowest exposure level with observable adverse effects. These endpoints are likely most common due to the lower amount needed of the toxicants in question and because they serve to delineate the threshold at which adverse events begin to occur.

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) #factor. So change it to a date
```

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

```
Litter$collectDate <- Litter$collectDate %>% as.Date(format = "%Y-%m-%d")
class(Litter$collectDate) #now it's a date.
```

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

```
#determine Aug. 2018 sampling dates:
Litter %>% filter(collectDate >= "2018-08-01" & collectDate <= "2018-08-31") %>% select(collectDate) %>%
```

```
##    collectDate
## 1   2018-08-02
## 92  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`?

```
#summary
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
```

```
#unique
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
```

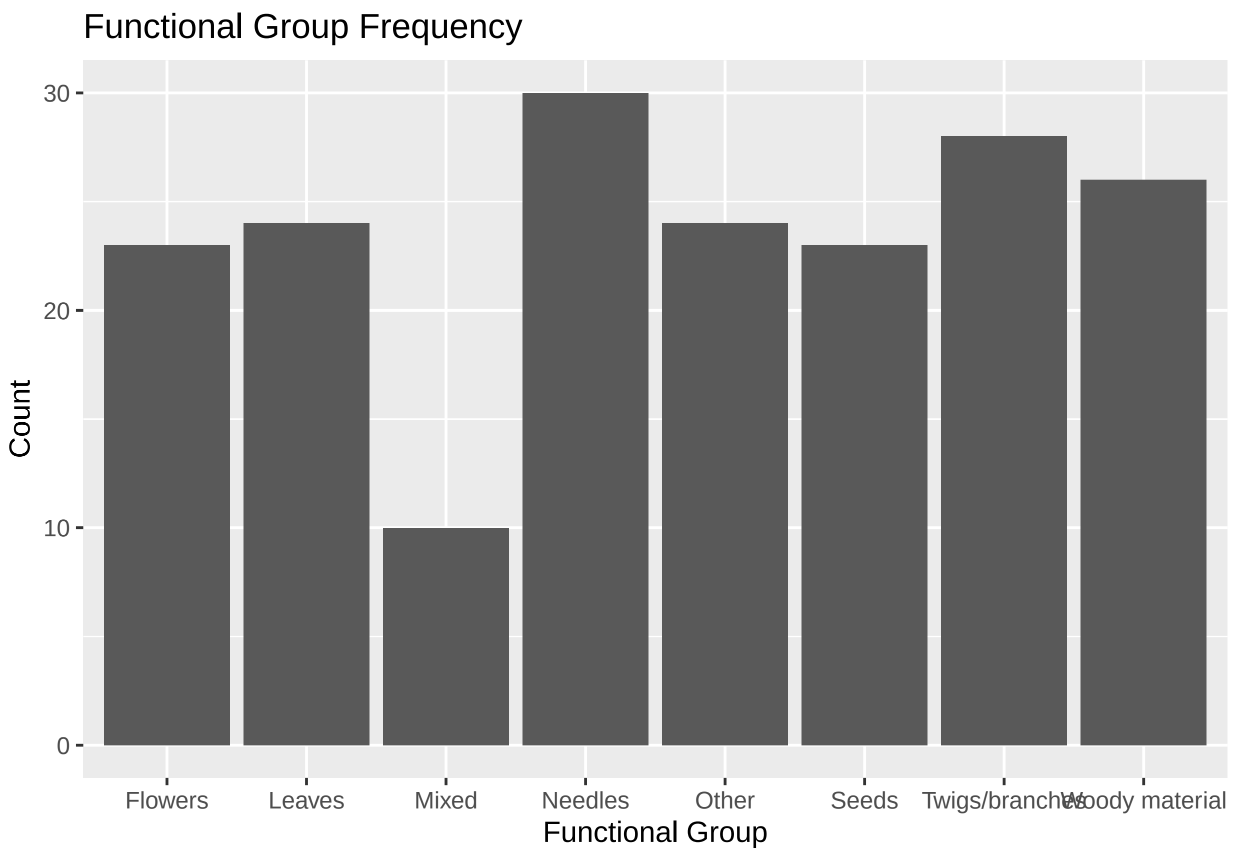
```
#number of plots
length(unique(Litter$plotID))
```

```
## [1] 12
```

> Answer: 'Summary' returns how many samples were taken at each plot at Niwot Ridge. 'Unique' simply returns each unique plot at which samples were taken (of which there are 12).

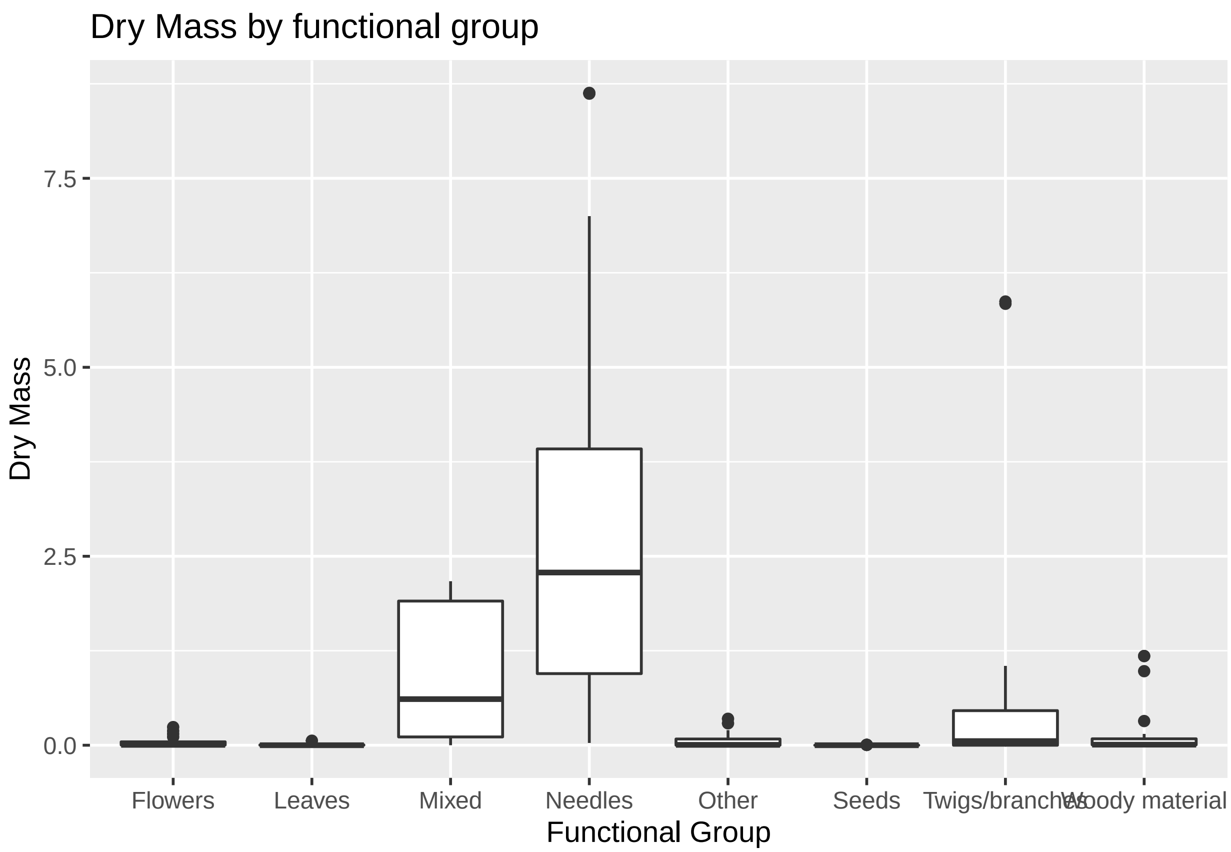
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
funcGroup_count <- ggplot(Litter) + geom_bar(aes(x = functionalGroup)) +
                                    labs(title = "Functional Group Frequency",
                                         x = "Functional Group",
                                         y = "Count")
funcGroup_count
```

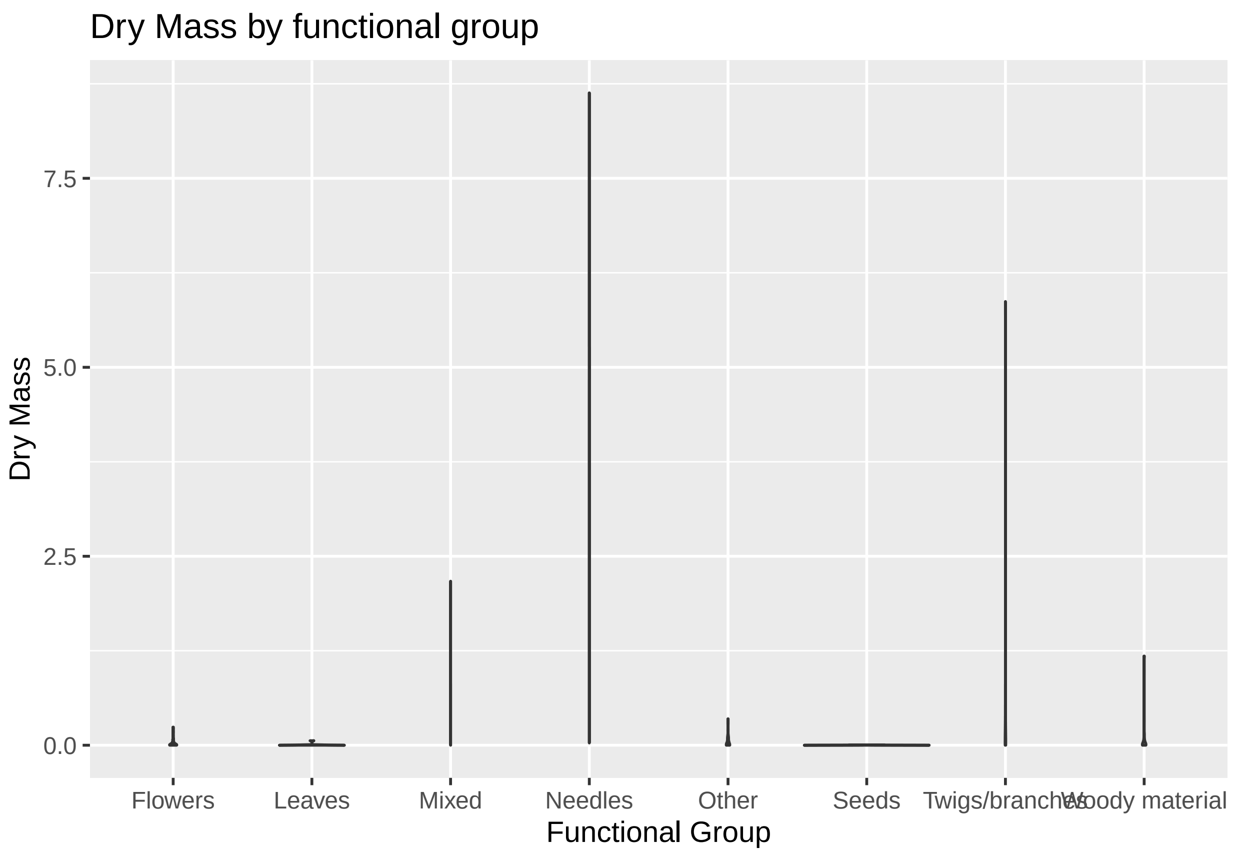


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
dryMass_box <- ggplot(Litter) + geom_boxplot(aes(x = functionalGroup, y = dryMass)) +
                                labs(title = "Dry Mass by functional group",
                                     x = "Functional Group",
                                     y = "Dry Mass")
dryMass_box
```



```
dryMass_violin <- ggplot(Litter) + geom_violin(aes(x = functionalGroup, y = dryMass)) +
                                   labs(title = "Dry Mass by functional group",
                                        x = "Functional Group",
                                        y = "Dry Mass")
dryMass_violin
```



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

> Answer: The data is not multimodal, so the boxplot's rectangular area is easier to view & understand than the lines on the violin plot.

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

> Answer: Needles, mixed, and twigs/branches/woody debris have the highest biomass at these sites.