Assignment 3: Data Exploration

Stephanie Kinser, Section #1

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 <>.

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.**

```
#install.packages("tidyverse")
#install.packages("ggforce")
library(tidyverse)
library(ggforce)


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 ecotoxicology of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

Answer: We may be interested in the ecotoxicology of neonicotinoids for several reasons. First, we may want to study the success of neonicotinoids in targeting certain insects that may be considered pests. While neonicotinoids may be designed to target some insects, we may also be interested in the indirect impacts on other insect species and how the chemicals are affecting insect populations more broadly. Finally, the ecotoxicology of neonicotinoids on insects may give clues to how the chemicals are affecting larger ecosystems.

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 may be interested in studying litter and woody debris to understand its effects on wildlife populations and ecosystem health. Litter and woody debris likely contributes to specific wildlife habitats which would affect species' populations and health. Litter and woody debris likely also have larger ecosystem benefits that affect forest health and wildlife.

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: *Samples are measured by calculating the mass of the litter and woody debris to an accuracy of 0.01 grams.* Ground trap samples are collected once a year; whereas elevated trap samples are collected more frequently. *The locations of traps may be randomized or targeted depending of the vegetation at the site.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) #dimensions of dataset
```

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

   Answer: The Neonics dataset has 4623 rows and 30 columns.

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)
```

```
##     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 and mortality are by far the most studied effects with over 1,000 entries each. Behavior, feeding behavior, and reproduction are also commonly studied, though far less so. Population and mortality may be of interest because they are indicators of ecosystem health and are fairly easy to measure.

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, 7) #used top 7 to account for Other which isn't a single species
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee      Italian Honeybee
##                   152                   140                   113
##               (Other)
##                  3083
```

Answer: The six most common insect species in the data set are: honey bee, parasitic wasp, buff tailed bumblebee, Carniolian honey bee, bumble bee, and Italian honey bee. Each of these species is a type of bee and they are important pollinators that support the growth of plant species, which then provide habitat and food for other species. Bees are an indicator species for the health of an ecosystem.

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.)
```

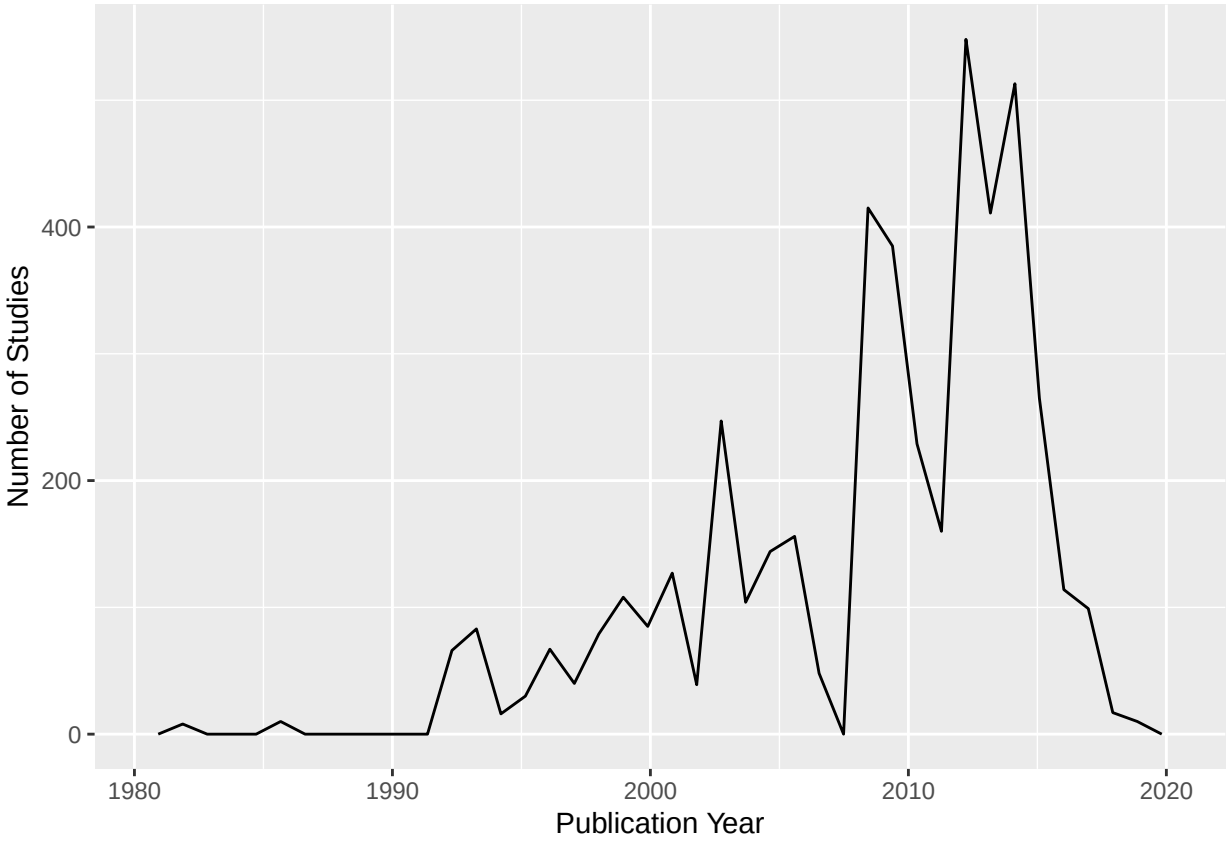
```
## [1] "factor"
```

Answer: The class is factor. Factors represent categorical data, meaning they assign a number value to represent a given category. A factor, therefore, is not numeric because performing mathematical operations on a factor would be meaningless.

Explore your data graphically (Neonics)

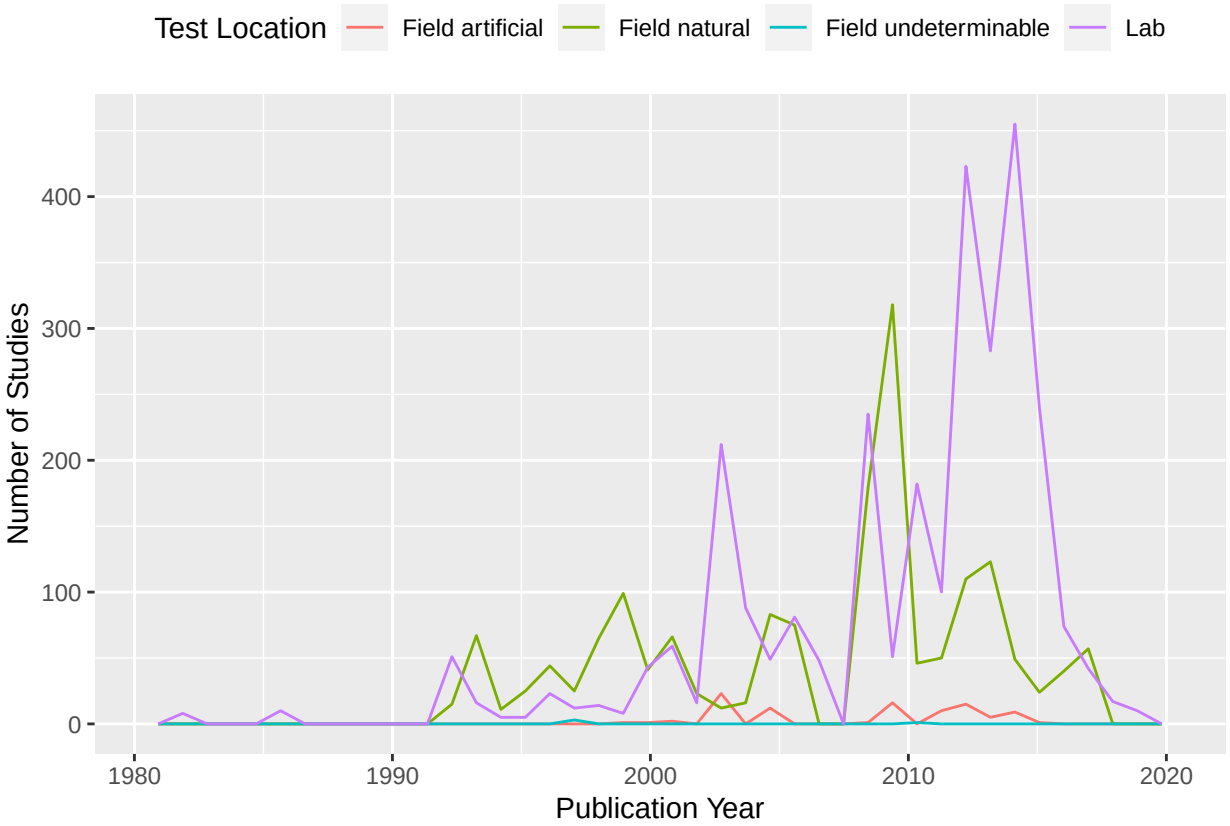
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics)+
geom_freqpoly(aes(x = Publication.Year), bins = 40)+
  labs(x = "Publication Year", y = "Number of Studies")
```



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 = 40)+
  theme(legend.position = "top")+
  labs(x = "Publication Year", y = "Number of Studies", color = "Test Location")
```

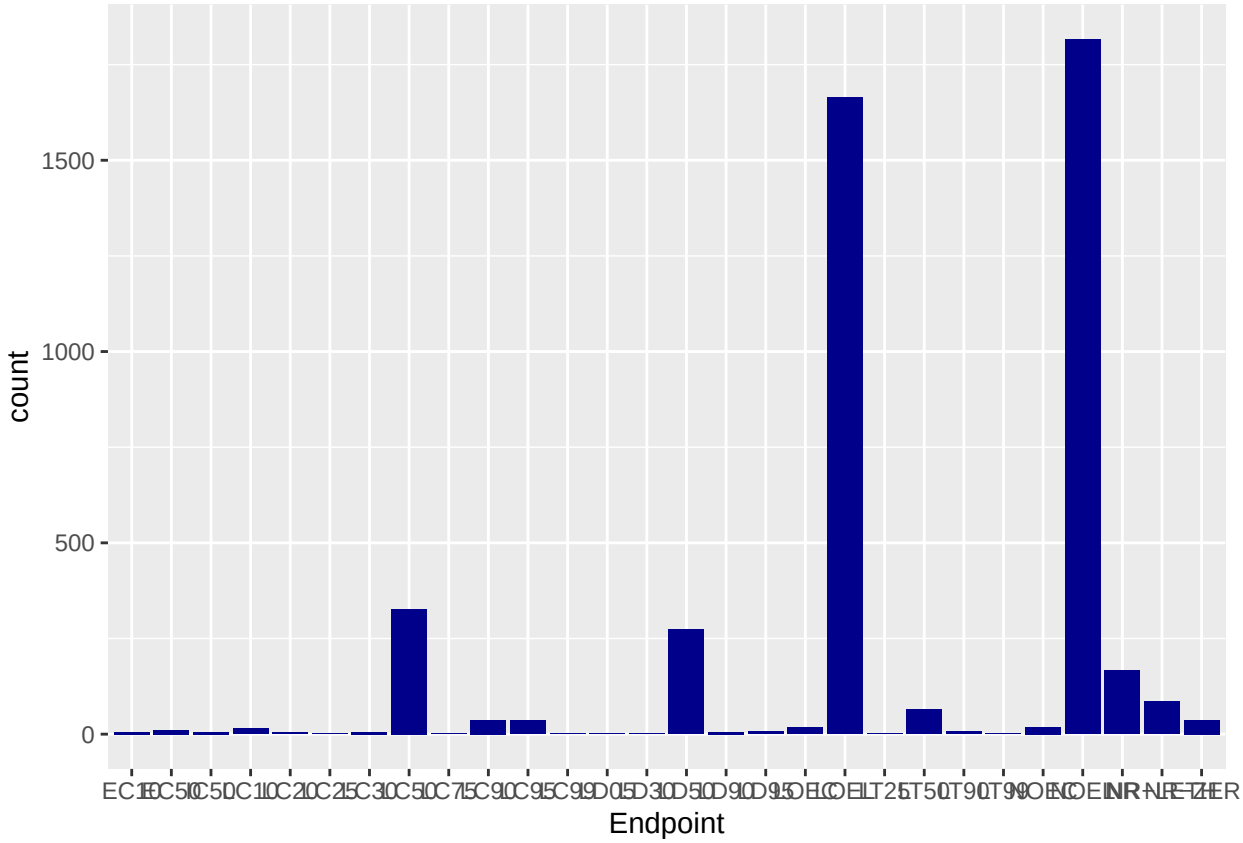


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: A lab is the most common test location during the 1980s and early 1990s. In the mid to late 1990s, a natural field becomes the most common test location. After the 2000s, the lab once again becomes more common with a natural field peaking around 2009. Artificial and undeterminable fields are never the most common test site.

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(fill = "darkblue")
```



Answer: The two most common endpoints are NOEL and LOEL. NOEL stands for no-observable-effect-level, meaning that the highest concentrations of the chemical does not show significantly different effects from a control. On the other hand, LOEL stands for lowest-observable-effect-level, meaning that the lowest concentration produced statistically significant effects that were different from the 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] "factor"
```

```
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d") #change from factor to date clas
class(Litter$collectDate)
```

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

```
unique(Litter$collectDate) #determine sample dates
```

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

Answer: Litter was samples on August 2nd and August 3rd 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`?

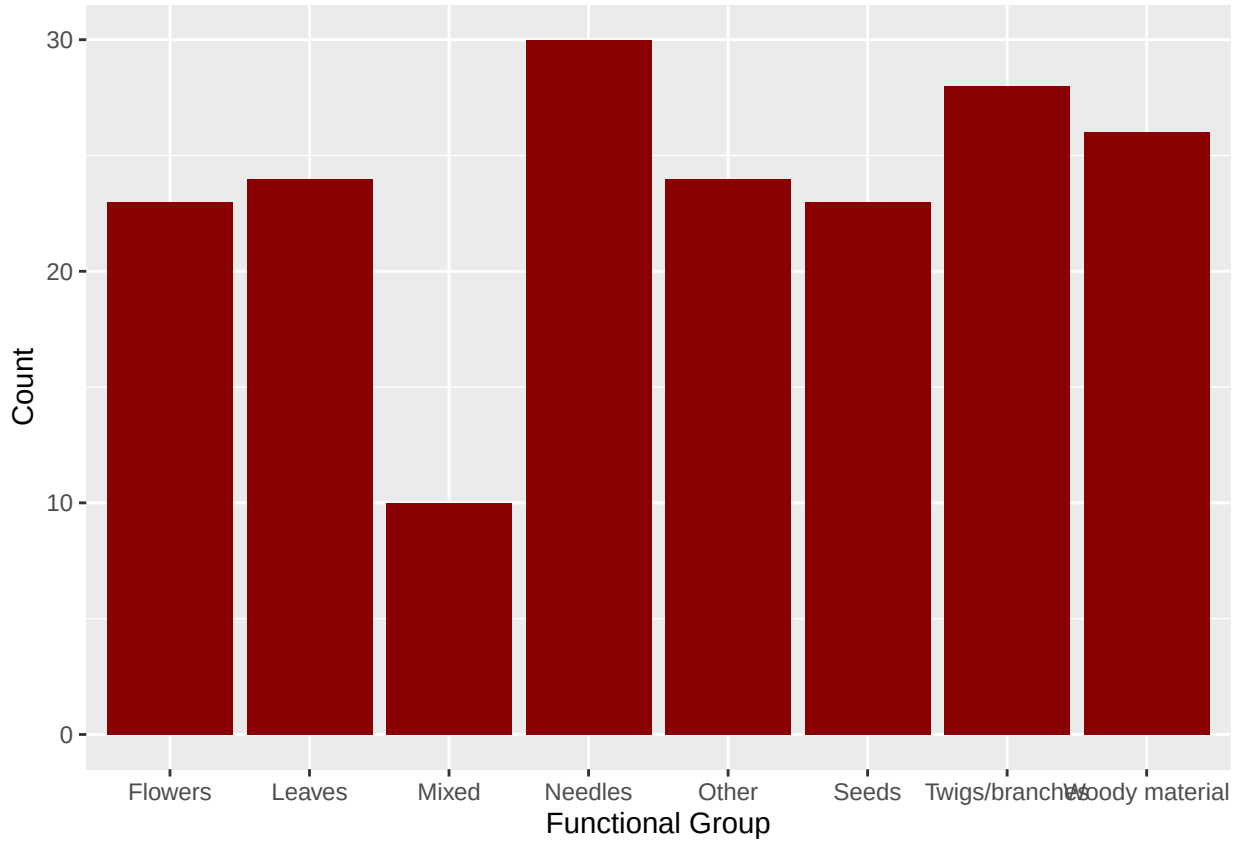
```
unique(Litter$plotID) #determine # of plots sampled
```

```
##  [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 unique plots were sampled. The unique function returns each unique element for a given vector in the dataset. The summary function returns the number of each element for a given vector in the dataset.

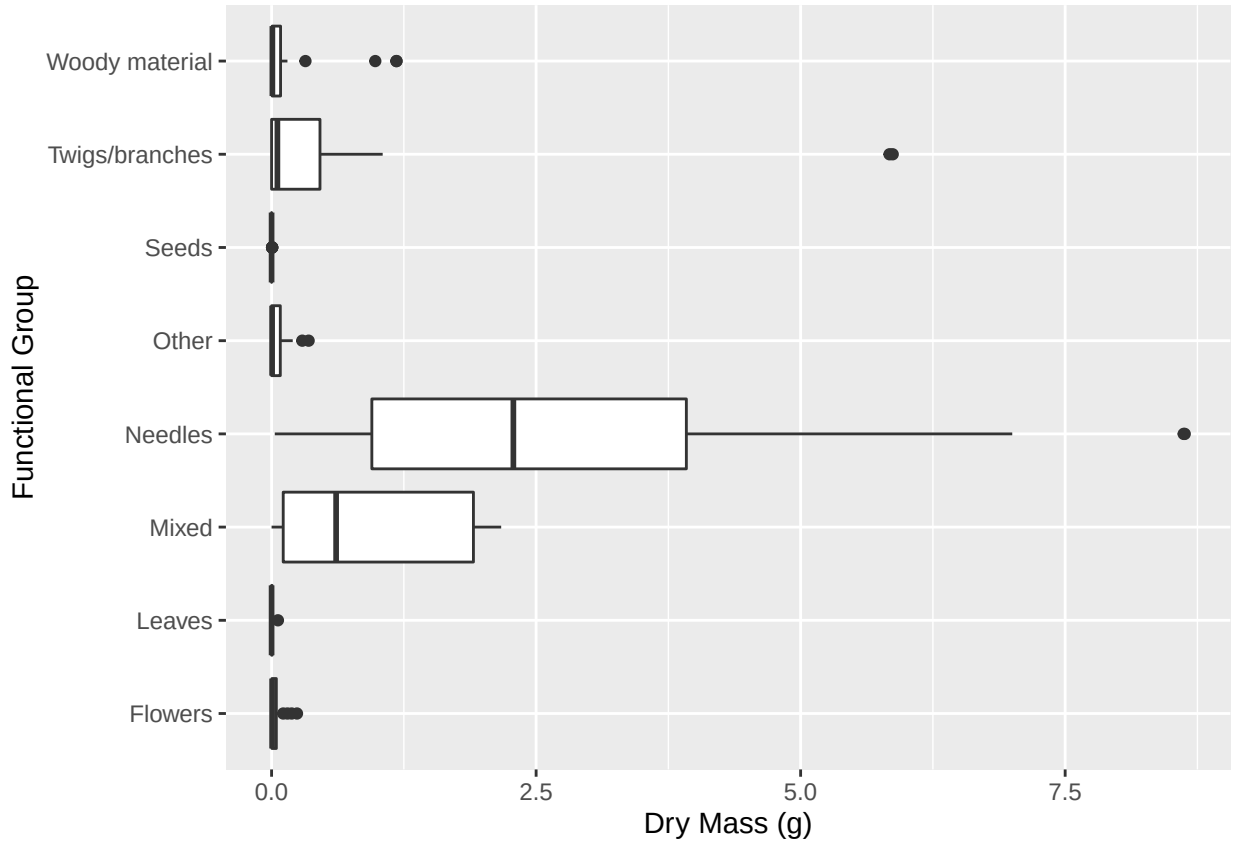
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(x = functionalGroup))+
         geom_bar(fill = "darkred")+
  labs(x = "Functional Group", y = "Count")
```

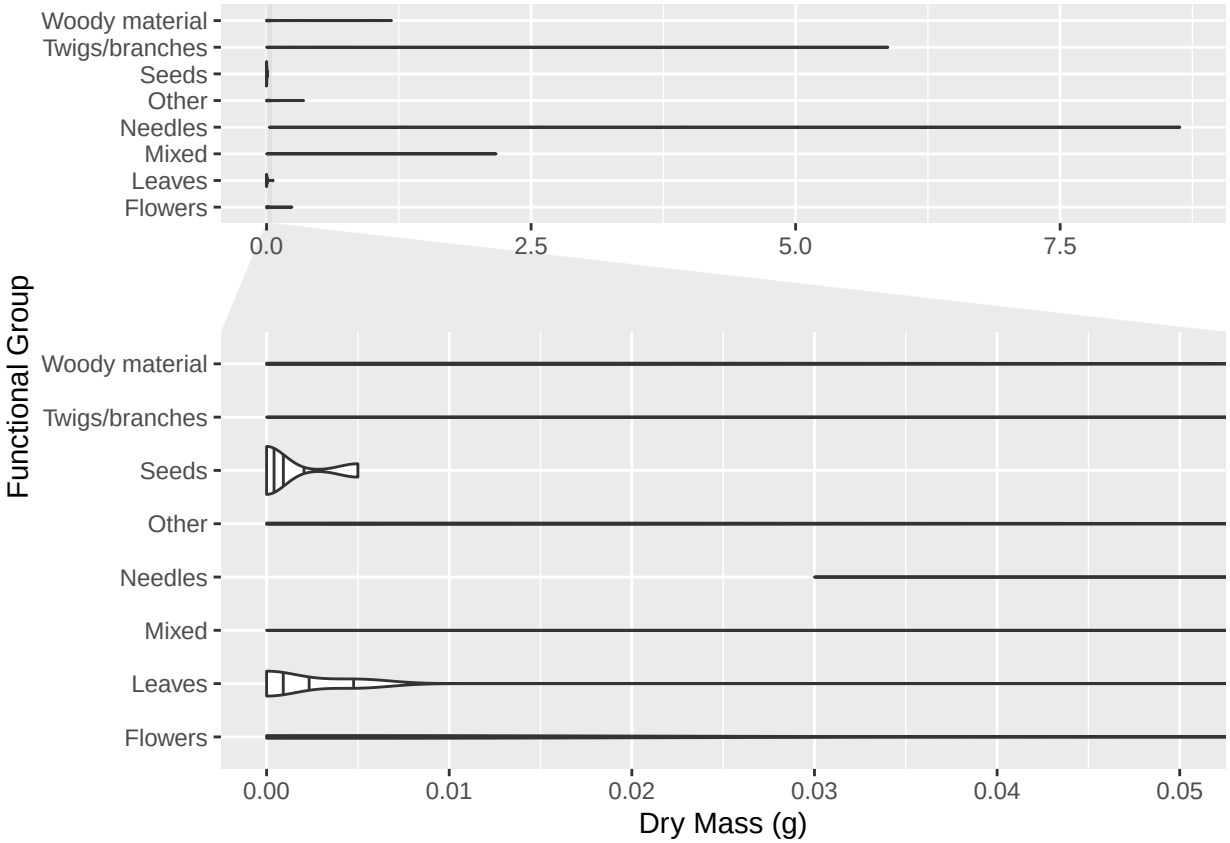


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(Litter) +
  geom_boxplot(aes(x = dryMass, y = functionalGroup))+
  labs(x = "Dry Mass (g)", y = "Functional Group")
```




```
ggplot(Litter) +
  geom_violin(aes(x = dryMass, y = functionalGroup), #original violin plot
              draw_quantiles = c(0.25, 0.5, 0.75))+
  facet_zoom(xlim = c(0, 0.05))+ #zoomed in range to see violin shape
  labs(x = "Dry Mass (g)", y = "Functional Group")
```



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

> Answer: The boxplot is a much more effective visualization here because the mass range is so large that the violin plot is too difficult to see. Because the box plot separates outliers as points, it is easier to understand the range of values for Dry Mass.

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

> Answer: Needles tend to have the highest biomass at the sites.