Assignment 3: Data Exploration

Laura Martinez

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. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.
6. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai.

The completed exercise is due on Sept 30th.

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 include the subcommand to read strings in as factors.

```
setwd("~/Documents/EDA-Fall2022")
getwd()
```

```
## [1] "/Users/laura/Documents/EDA-Fall2022"
```

```
# install.packages(tidyverse)
library(tidyverse)
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: The EPA put together a comprehensive knowledgebase of known chemicals adverse health affects on aquatic and terrestrial species, which is especially important for assessing the impact of chemicals on biota. Understanding the impact of chemicals used in the agricultural sector such as Neonicotinoids is important for determining the uintended negative side effects on an entire food web.

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: Natural processes, such as tree fall and decay, are integral to the balance of natural systems like Niwot Ridge in Colordo. Tree fall and decay contribute many nutirents to the forest floor, and depending on the size and location of the debri they confer different purposes to the forest floor. Larger debris takes longer to break down and release nutrients, and trees that are closer to the forest floor will retain more mositure thus decaying more rapidly (https://placebasedbasics.weebly.com/fine-and-course-woody-debris-analysis.html).

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. Matter was collected from NEON sites and sorted by size to determine matter type. Litter is defined as matter that has dropped from the forest canopy, and is at most 2 cm in diameter and less than 50cm, whereas fine wood debris can be at most 2 cm in diameter, but is longer than 50cm. 2. For the spatial sampling design "Trap placement within plots may be either targeted or randomized, depending on the vegetation." 3. Interestingly, the temporal sampling design is also highly variable across grids."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?

```
# number of rows in Neonics
nrow(Neonics)
```

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

```
# number of columns in Neonics
ncol(Neonics)
```

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

```
# number of rows by number of columns in Neonics
dim(Neonics)
```

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

```
# length of objects in the dataset Neonics
length(Neonics)
```

```
## [1] 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)
```

```
##     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: Based on the number of entries for the effects studied, Mortality and Population are the most commonly studied, with mortality having 1,493 hits and population having 1,803 hits. This is because Neonicotinoids have an immediate and adverse impact on inverterbrates's fatality rates and population numbers.

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.

```
# Create new data frame with species count
Species <- as.data.frame(summary(Neonics$Species.Common.Name))

# Use head function to determine top 6 results
head(Species)
```

```
##                       summary(Neonics$Species.Common.Name)
## Honey Bee                                              667
## Parasitic Wasp                                         285
## Buff Tailed Bumblebee                                  183
## Carniolan Honey Bee                                    152
## Bumble Bee                                             140
## Italian Honeybee                                       113
```

Answer: 5 of the 6 top species are bees, and all 6 species are important pollinators. They are key in the dispersal of pollen for plants.

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.)
```

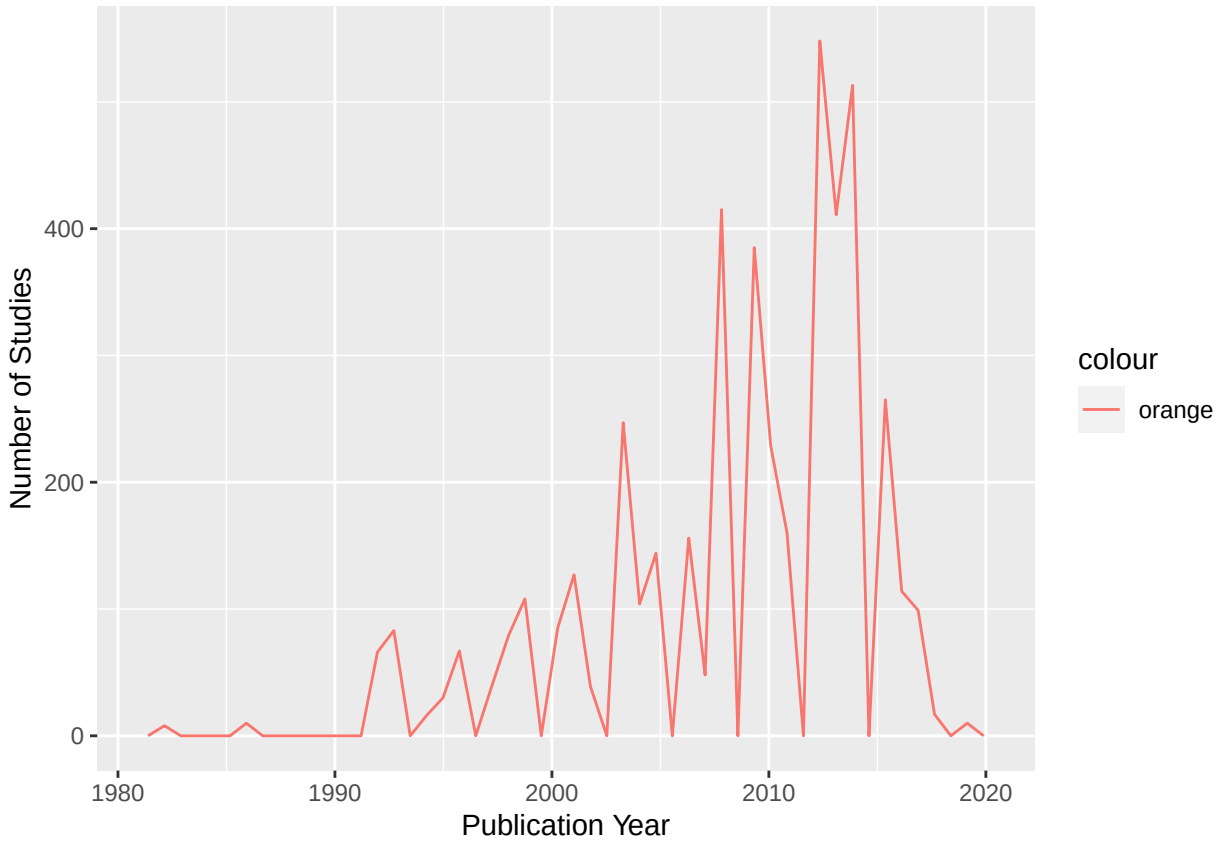
```
## [1] "factor"
```

Answer: Conc.1..Author is classified as a factor. It is classified as a factor and not numeric, because the values have addiitional characters such as "/" and equality statements. A formatting function would be required to make this a numeric dataset that recognized the numeric values versus the characters.

Explore your data graphically (Neonics)

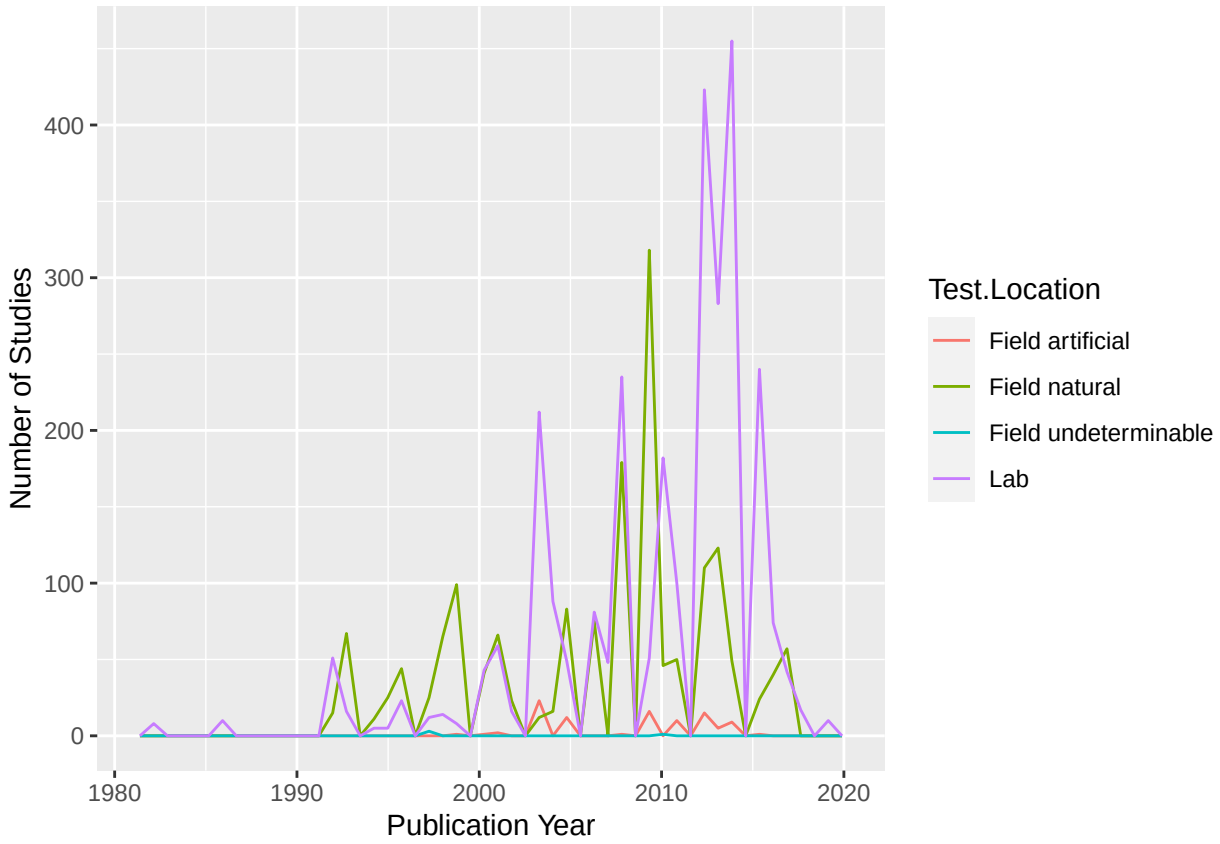
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year, color = "orange"), bins = 50) +
    xlab("Publication Year") + ylab("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 = 50) + xlab("Publication Year") + ylab("Number of Studies")
```

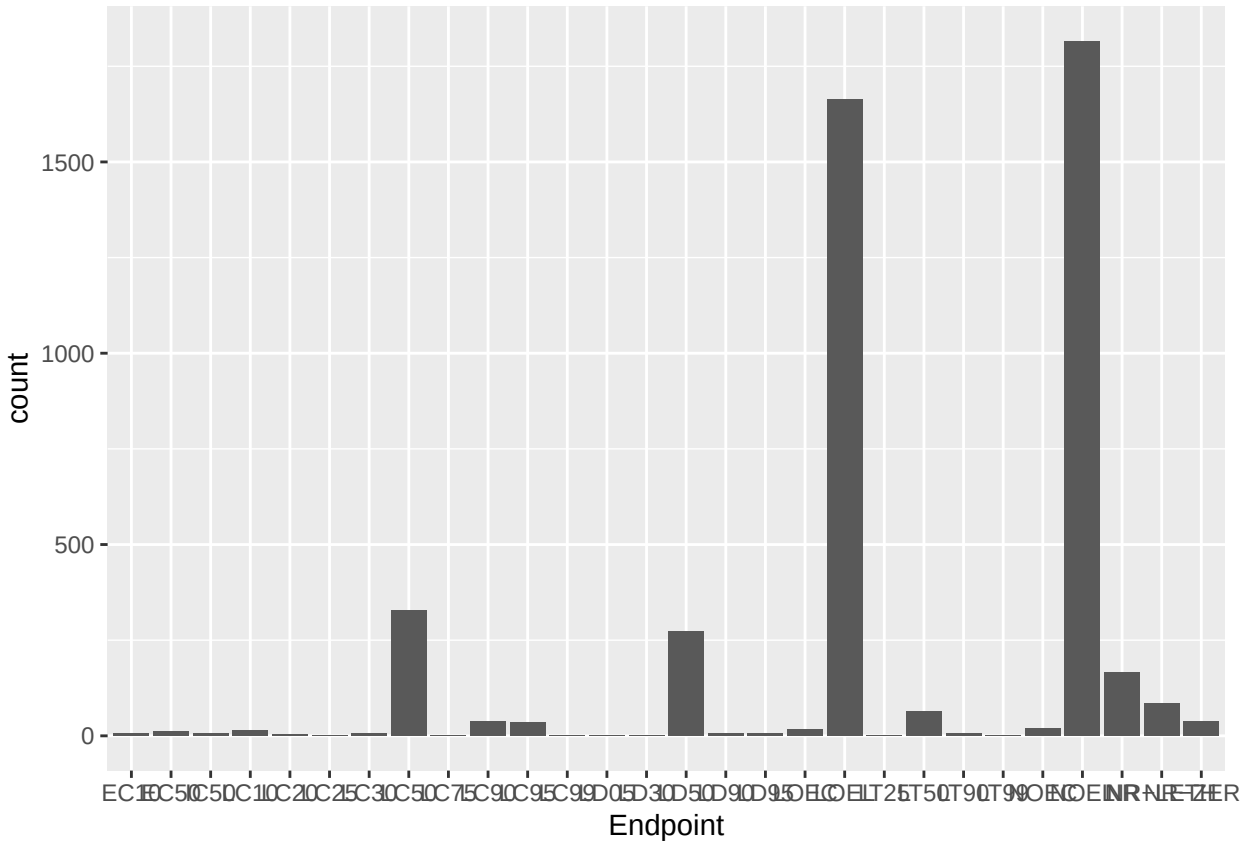


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: According to the frequency polygon, natural fields (green) and artificial fields (red) are the most commonly sampled test locations. After 2000, there was a drastic surge in the use of lab fields for data collection. Around the 2010s, this shifted to natural fields being the more popular choice for test locations until a few years later when dominance shifted back to lab test locations.

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) + geom_bar(aes(x = Endpoint))
```



Answer: The two most common end points are LOEL and NOEL, both having counts over 1,500. LOEL is used for terrestrial settings and is defined as the Lowest-Observable-Effect-Level, which is the lowest dose at which observable differences can be seen. NOEL is also used in terrestrial settings and is defined as No-Observable-Effect-Level, which is the highest dose that does not produce results different from reported levels from statistical tests.

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.

```
# collectDate class
class(Litter$collectDate)
```

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

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

# collectDate reclassification
class(Litter$collectDate)
```

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

```
# Litter sampled in Aug 2018
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`?

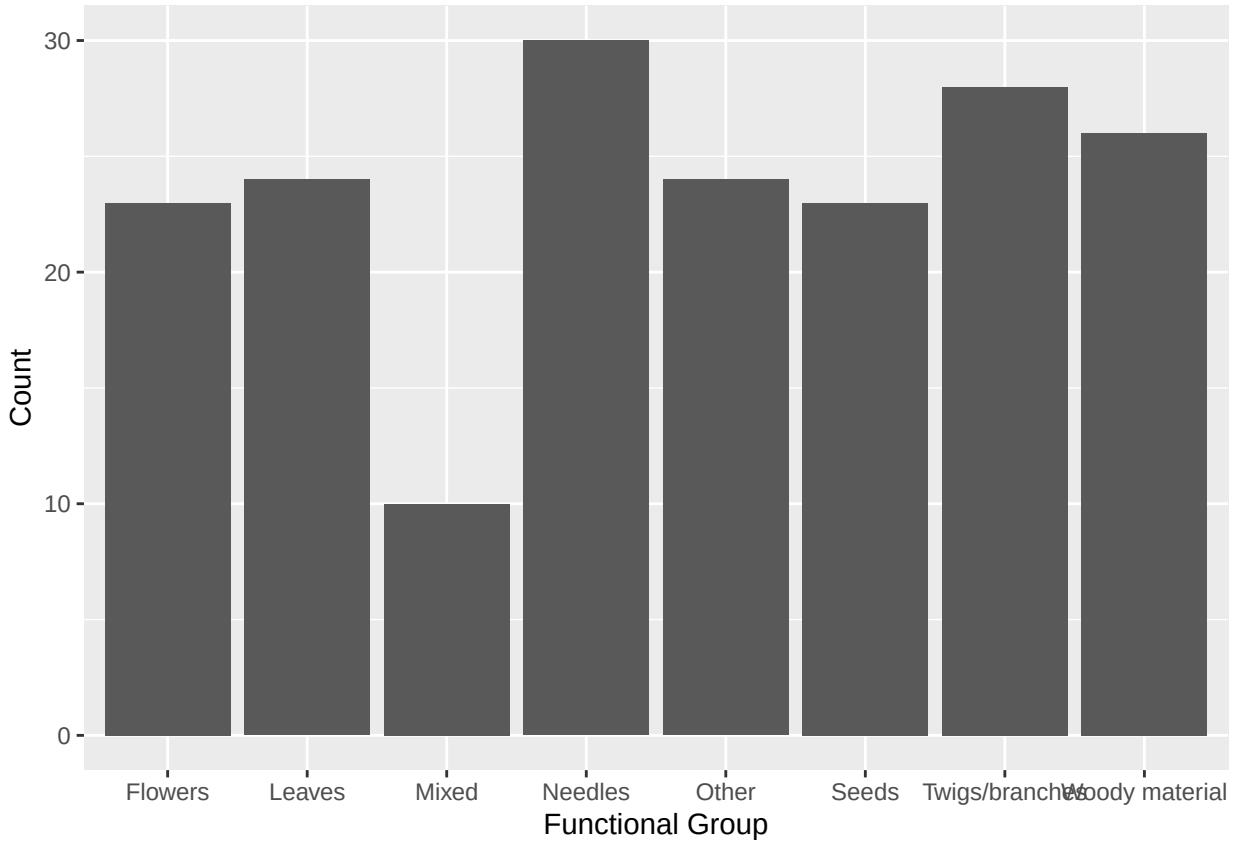
```
unique(Litter$plotID)
```

```
##  [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: The unique functions returns the 12 plots sampled. This is different from a summary function that would return the number of times/count for each plot sampled.

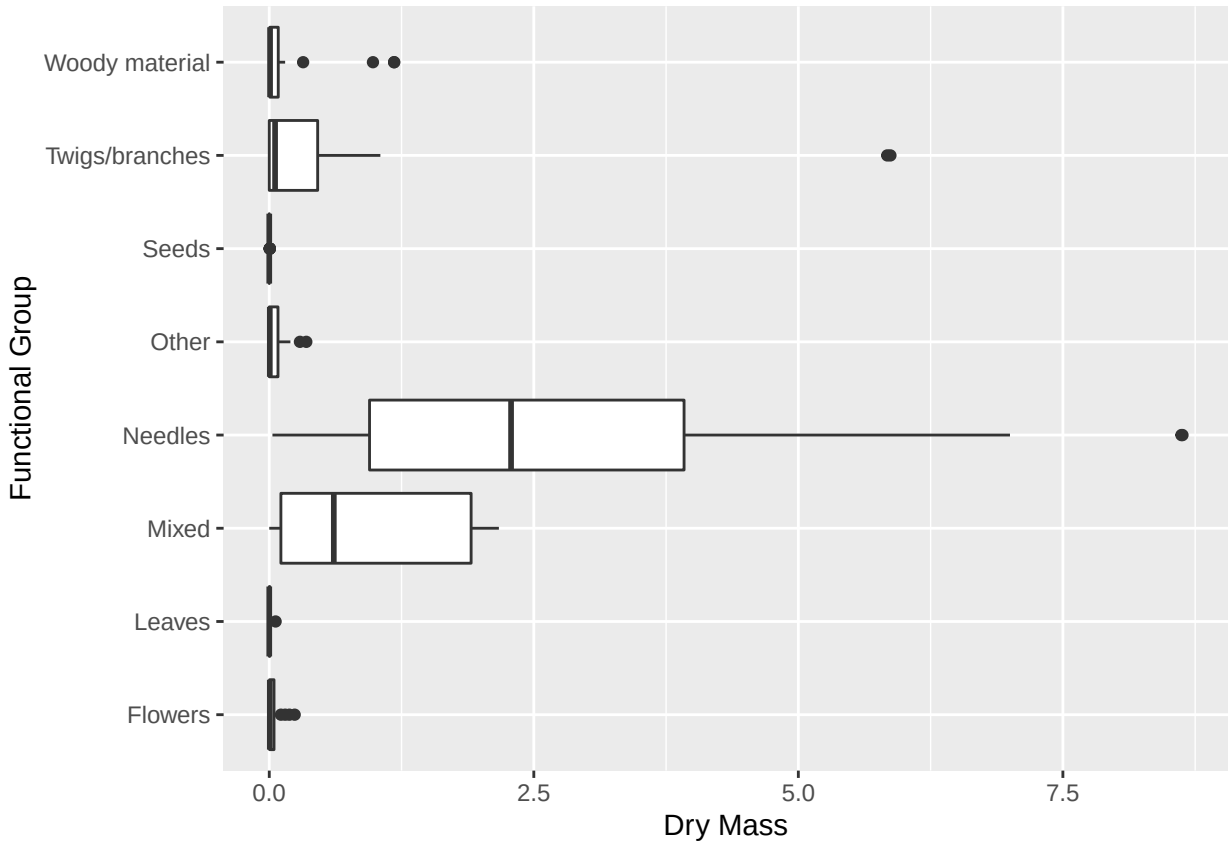
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter) + geom_bar(aes(x = functionalGroup)) + xlab("Functional Group") +
    ylab("Count")
```

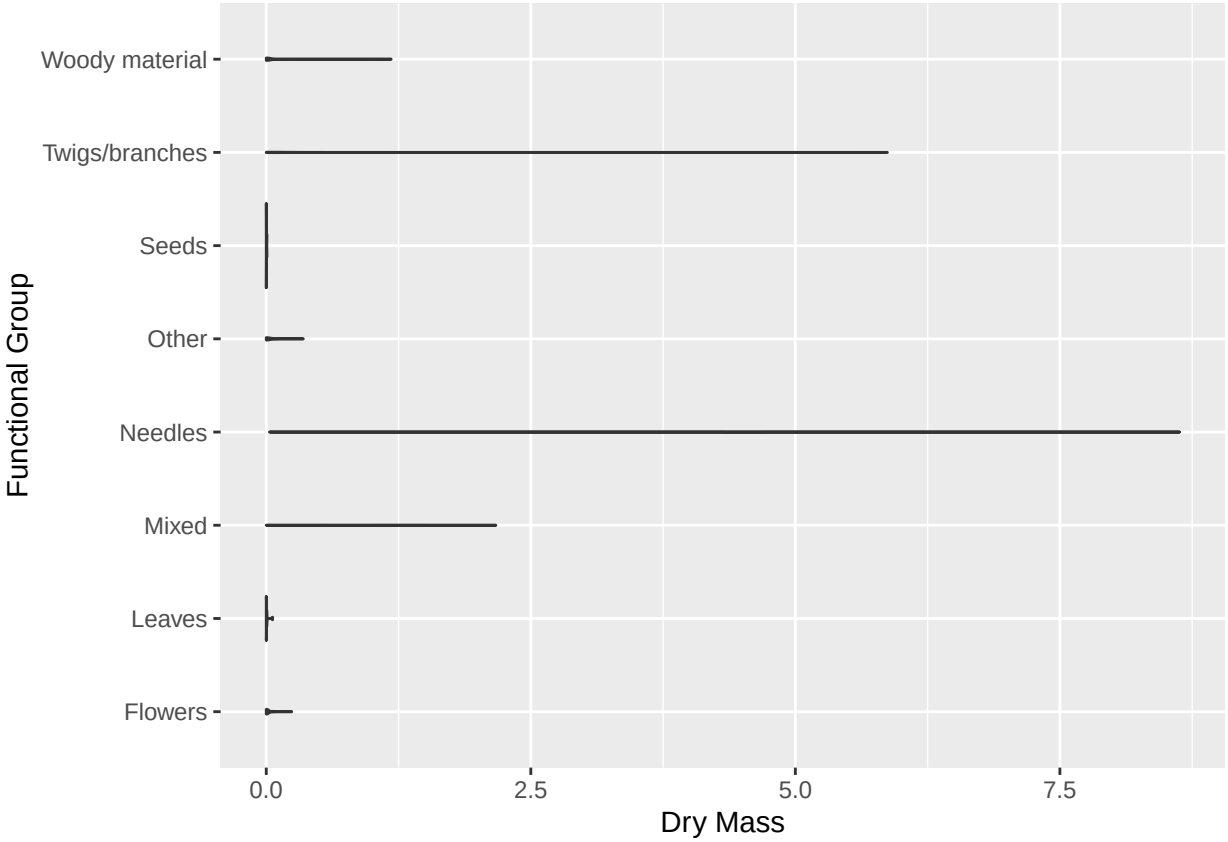


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
# Box Plot
ggplot(Litter) + geom_boxplot(aes(x = dryMass, y = functionalGroup)) + xlab("Dry Mass") +
    ylab("Functional Group")
```




```
# Violin Plot
ggplot(Litter) + geom_violin(aes(x = dryMass, y = functionalGroup)) + xlab("Dry Mass") +
    ylab("Functional Group")
```



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

> Answer: The boxplot visualization includes a lot more helpful information to understanding the data, such as the mean, the interquartile ranges and the comparison of IQR across functional groups, and outlier that could skew the data.

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

> Answer: Needles and Mixed types of litter tend to have the highest biomass based on the mean mass recorded acorss functional groups.