Assignment 3: Data Exploration

Justin Maynard

Fall 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.

```
getwd()
```

```
## [1] "/Users/justinmaynard/Fall_2023_EDE"
```

```
library(tidyverse)
library(lubridate)
library(ggplot2)
library(lubridate)
#Loaded libraries
```

```
Neonics <- read.csv("/Users/justinmaynard/Fall_2023_EDE/Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"
Litter <- read.csv("/Users/justinmaynard/Fall_2023_EDE/Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.c
#Imported data to 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 ecotoxicology of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: Neonicotinoids (Neonics) are the most widely used class of insecticide, and they work through affecting the central nervous system of insects. Because of this mode of attack, neonics can affect target and non target insects. Additionally, they are water soluble, and developing plants can absorb the neonics. Because neonics spread throughout an environment quickly, non target insects such as bees are at risk of death.

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 on the forest floor can be representative of the forest health and forest composition.

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. Sampling occurs only in tower plots, of which the locations were selcted randomly within 9-% flux footprint of the pirmary and secondary airsheds. 2. Litter sampling is targeted to takeplace in 20 40m x 40m plots in sites with forested tower airsheds. In sites with low saturated vegetation over the tower airsheds the litter sampling occured in 4 40m x 40m plots and 26 20m x 20m plots. 3. Plot centers must be more than 50m from paved roads and buildigns, plot edges must be 10m from dirt roads, and streams larger than 1m may not intersect the plots.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
print(dim(Neonics))
```

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

```
#Check dimensions of Neonics
```

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?

```
print(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
```

```
#Print summary of Neonics Effect
```

Answer: These effects may be of interest as they mostly are associated with negative effects on insects (mortality, intoxication, feeding behavior, etc). Knowing which effects are most common is a good place to start when thinking about which effects on insects are worth studying more in depth.

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...]

```
print(sort(summary(Neonics$Species.Common.Name), decreasing = TRUE)[0:7])
```

```
##               (Other)             Honey Bee        Parasitic Wasp
##                   670                   667                   285
## Buff Tailed Bumblebee   Carniolan Honey Bee            Bumble Bee
##                   183                   152                   140
##      Italian Honeybee
##                   113
```

```
#Print summary of Neonics species, sort by descending and limit to 7 results (6 plus other category)
```

Answer: The most commonly studied species are bees or pollinators. These specifically are of interest over other insects as effects on pollinators have large implications for the ecosystem as a whole.

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?

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

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

```r
#Print class of Neoncis 'Conc.1..Author.'
```

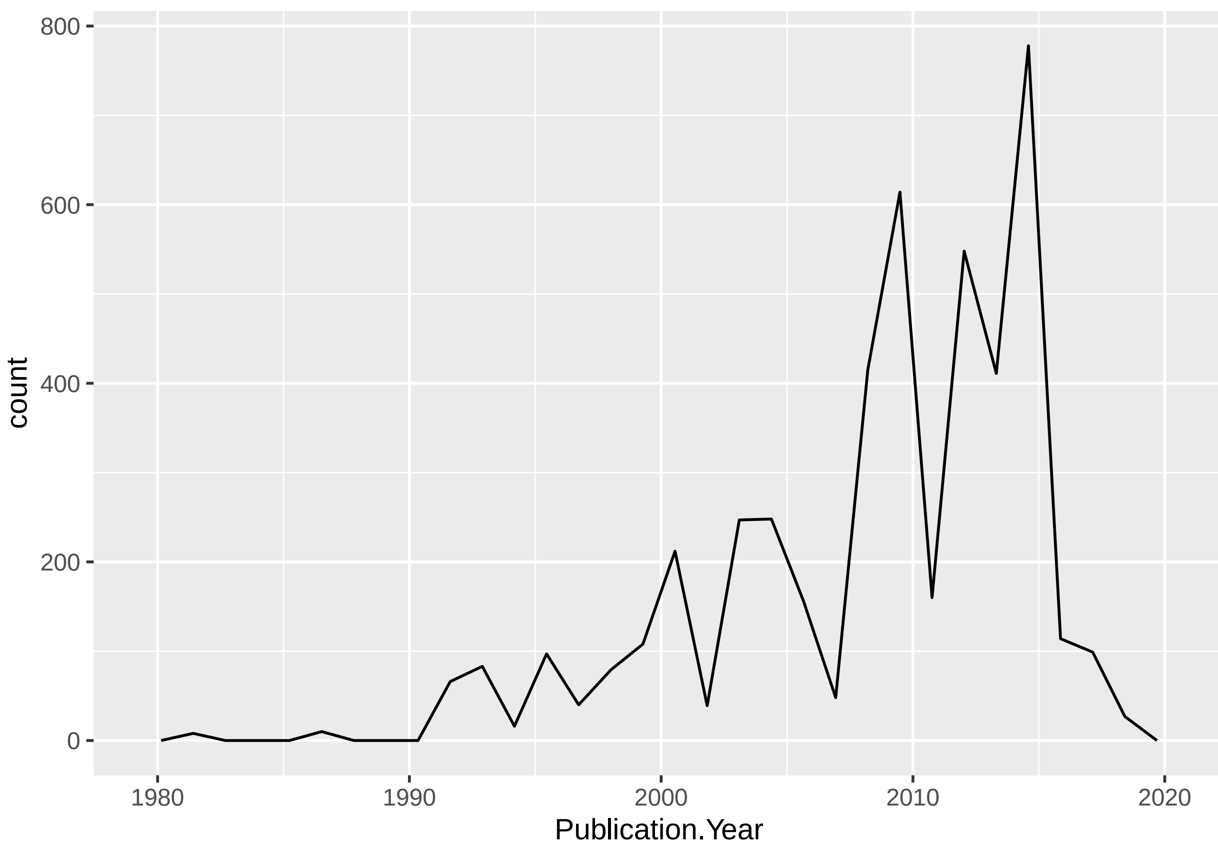
Answer: The datatype is a factor. This is such as it is a categorical variable and there are a set values that the data can take.

Explore your data graphically (Neonics)

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

```r
plot_q9 <- ggplot(Neonics) +
  geom_freqpoly(
    aes(x = Publication.Year))
print(plot_q9)
```

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

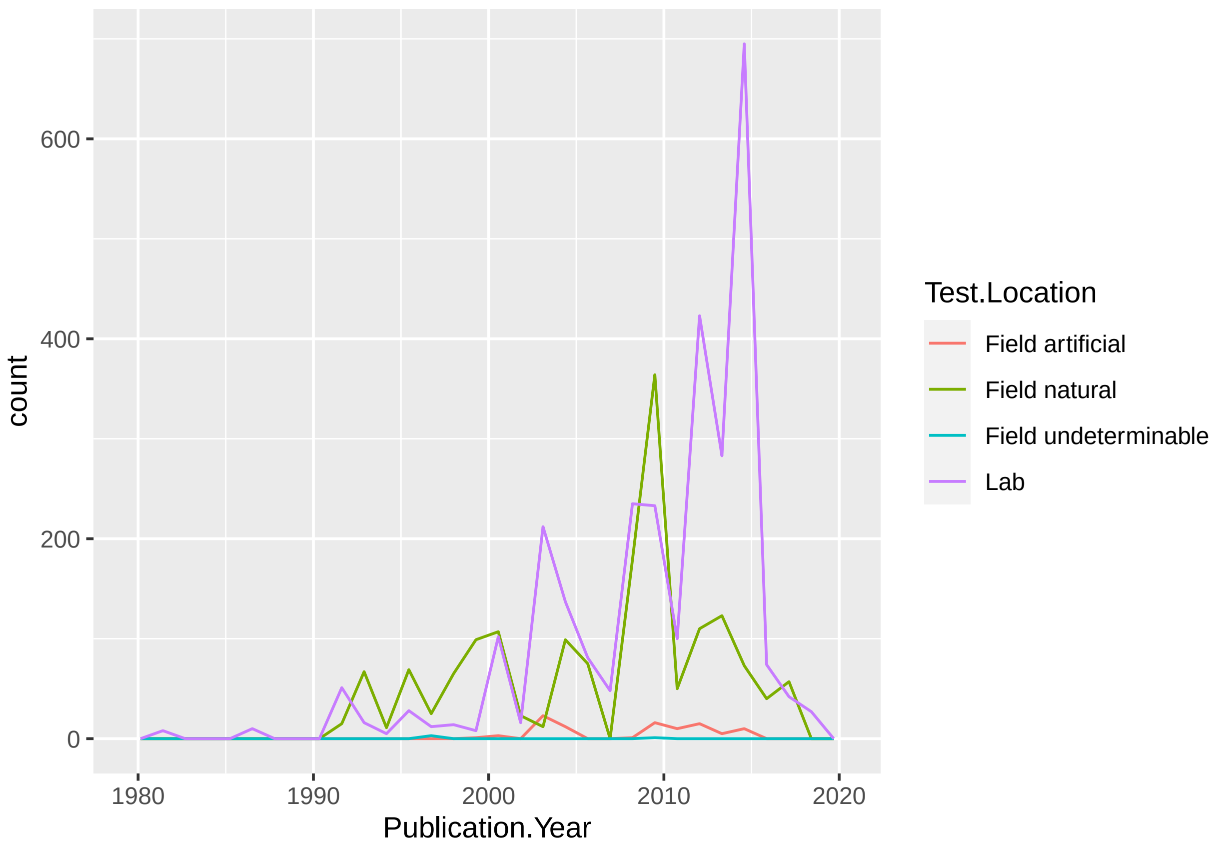



```r
#Create frequency plot of number of studies by publication year
```

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

```r
plot_q10 <- ggplot(Neonics) +
  geom_freqpoly(
    aes(x = Publication.Year, color = Test.Location))
print(plot_q10)
```

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




```r
#Create frequency plot of number of studies by publication year separated and colored by test location
```

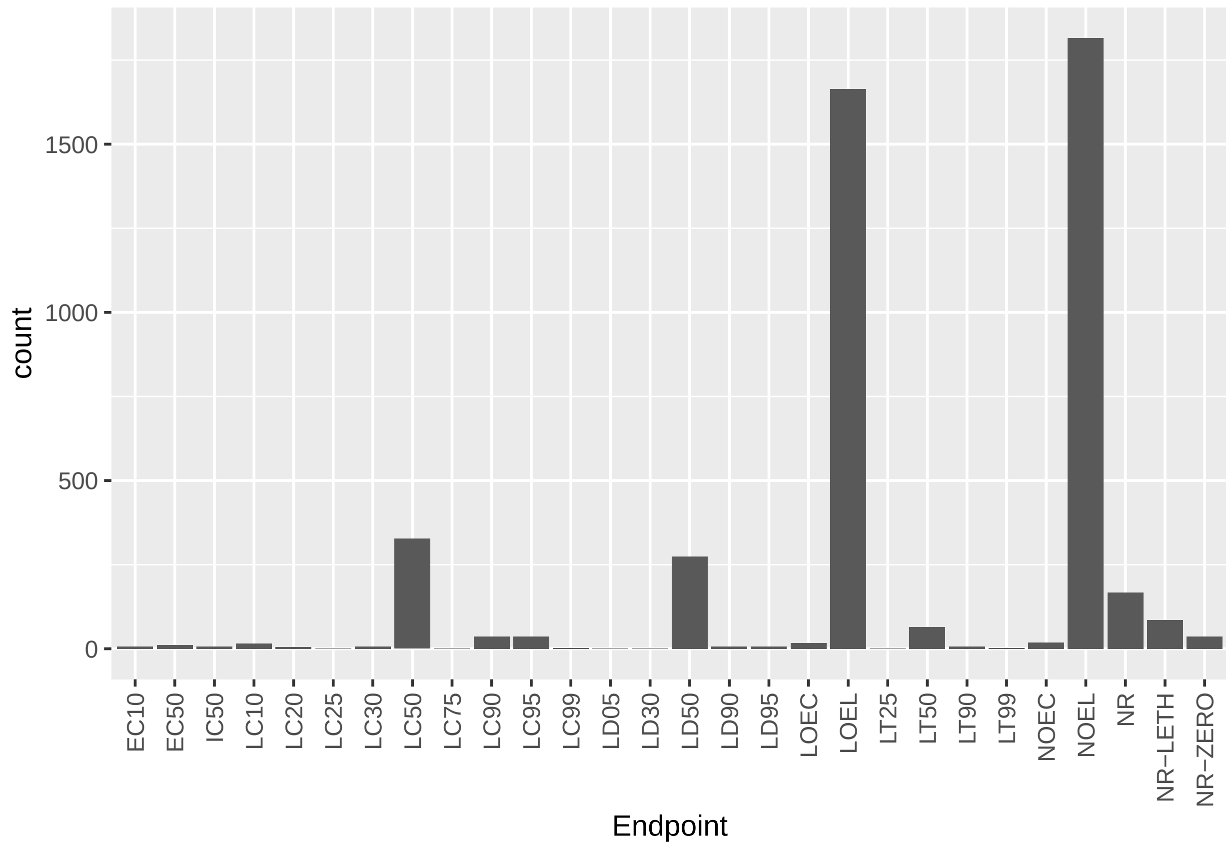
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test location is lab followed by field natural. Lab testing has varied over time, with it being less common than natural fields until around 2000. Additionally natural fields surpassed lab testing prior to 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.

[**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…]

```
plot_q11 <- ggplot(Neonics) +
       geom_bar(aes(x = Endpoint)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
print(plot_q11)
```




```
#Create bar graph of endpoint counts
```

> Answer: NOEL and LOEL are the two most common endpoints. LOEL is a terrestrial endpoint, and is defiend as the "Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC)." NOEL is also a terrestrial endpoint, and is defined as "No-observable-effect-level: highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test (NOEAL/NOEC)."

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.

```
print(class(Litter$collectDate)) #Print class of litter `collectData`
```

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

```r
Litter$collectDate2 <- ymd(Litter$collectDate) #Change collectDate to date using lubridate
print(class(Litter$collectDate2)) #Print new class of collectDate
```

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

```r
print(unique(Litter$collectDate2)) #Print the unique dates
```

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

```r
#Litter was sampled on August 2nd and 30th
```

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`?

```r
print(unique(Litter$plotID)) #Print unique values of 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
```

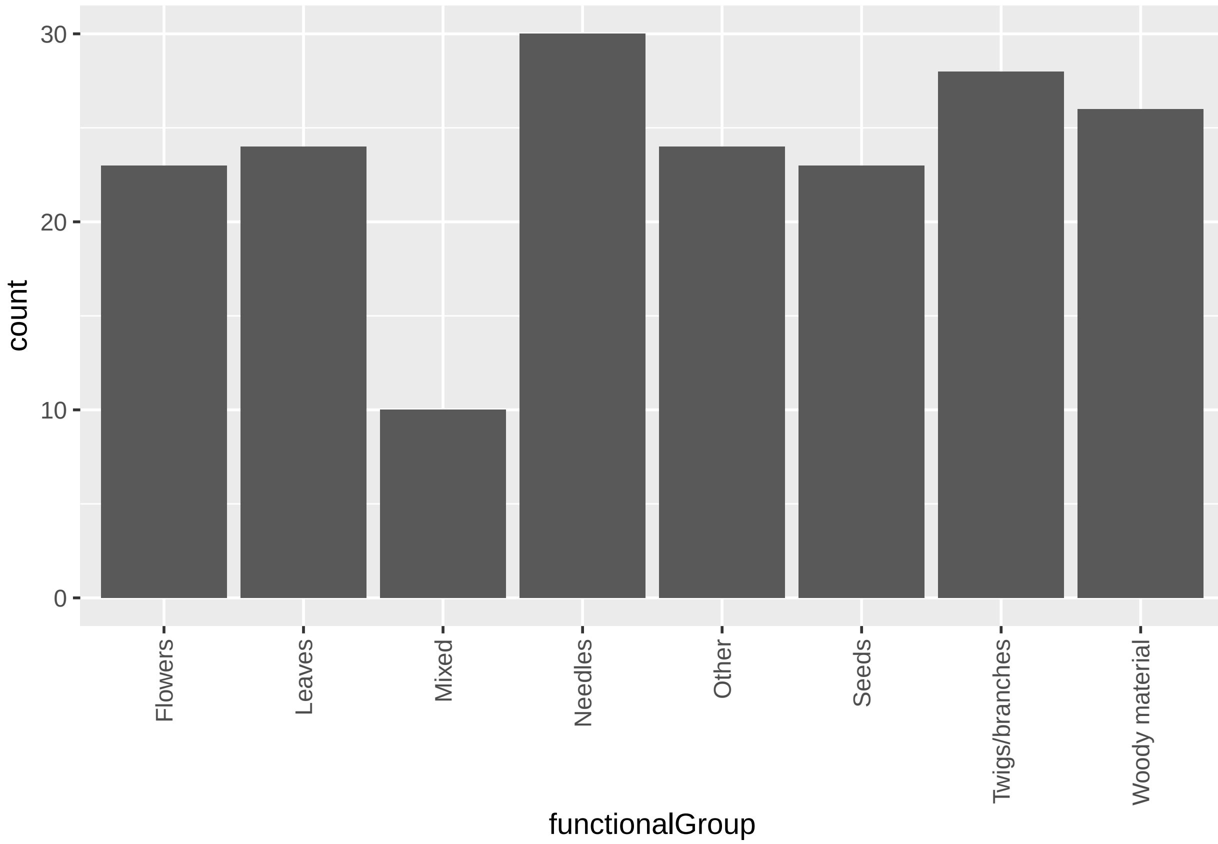
```r
print(summary(Litter$plotID)) #Print summary values of Litter 'plotID'
```

```
## 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
```

Answer: Unique tells you how many plots and what were sampled, summary tells you how many times each individual plot was sampled.

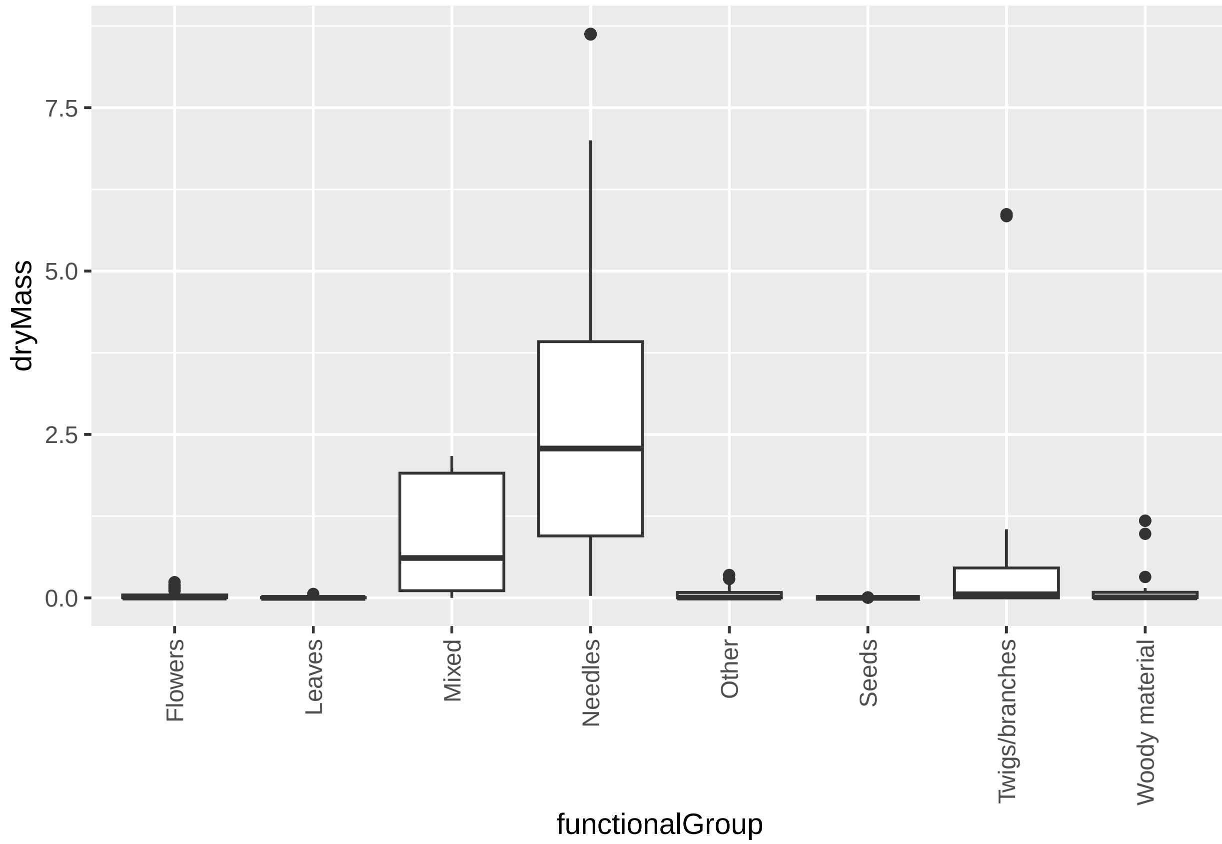
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
plot_q14 <- ggplot(Litter) +
       geom_bar(aes(x = functionalGroup))  +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
#Create bar chart of functionalGroup by count
print(plot_q14)
```

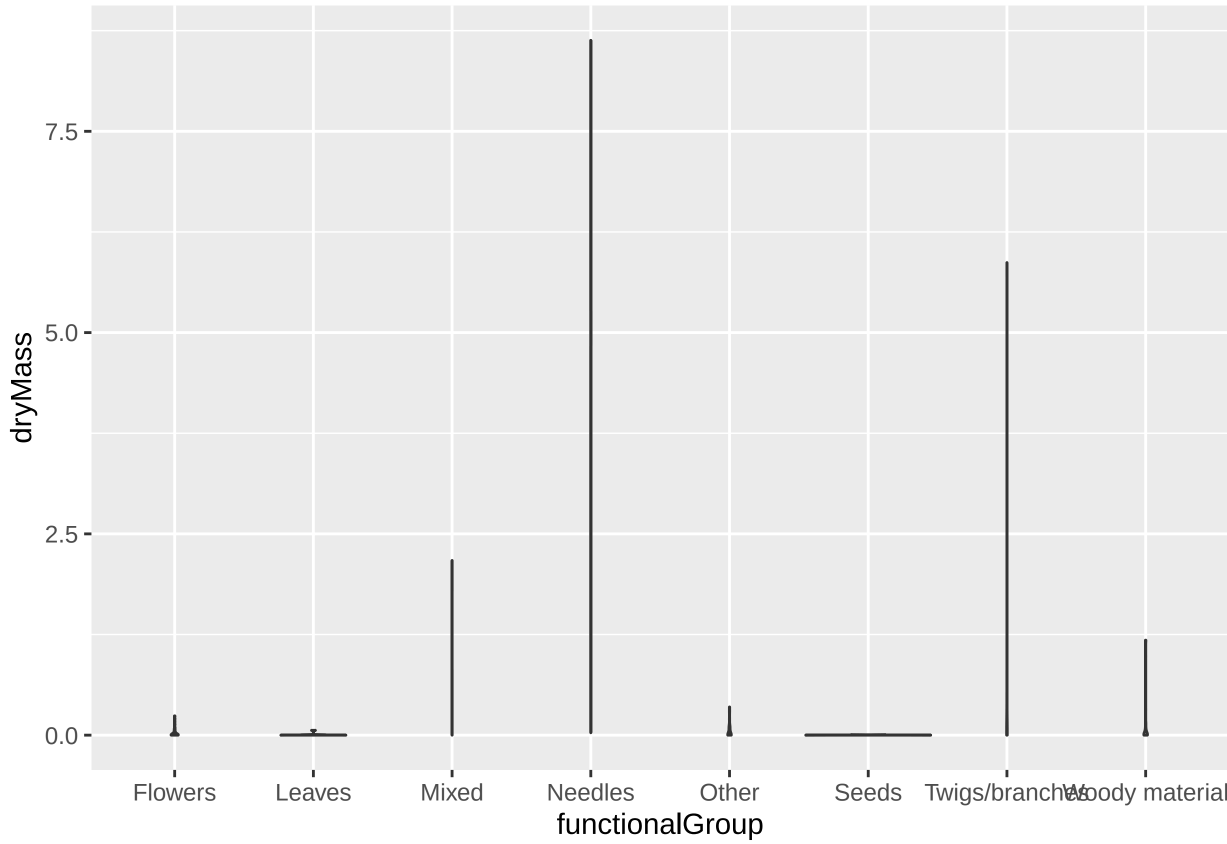


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
plot_15_a <- ggplot(Litter) +
       geom_boxplot(aes(x = functionalGroup, y = dryMass))  +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
#Create boxplot of functionalGroup by dryMass
print(plot_15_a)
```




```r
plot_15_b <- ggplot(Litter) +
  geom_violin(aes(y = dryMass, x = functionalGroup))
#Create violin plot of functionalGroup by dryMass
print(plot_15_b)
```



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

> Answer: Both plots allow you to see the dry mass by functional group, but the boxplot contains the whisker charts that allow you to see the quartiles, median, and outliers in the data. Additionally, the box plot is more visually appealing as the violin can be difficult ot read.

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

> Answer: Needles