Assignment 3: Data Exploration

Lauren Ng

Fall 2024

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 Canvas.

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. Load necessary packages (tidyverse, lubridate, here), check your current working directory 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.

```
#Load packages: tidyverse, libridate, here
library(tidyverse)
library (lubridate)
library (here)

#Use getwd() to reveal what R sees as the working directory
getwd()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```r
#Use here to show where the R Project file is
here()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```r
Neonics <- read.csv(
  file = here("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"),
  stringsAsFactors = TRUE)

Litter <- read.csv(
  file = here("./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 might want to understand how specific neonicotinoids are in their toxicity to insects we want to eliminate (pests) versus insects that we do not want to harm. We may need to see at what concentrations neonicotinoids are lethal to certain insects so that we can minimize broader damage to the ecosystem.

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 and litter serve many important functions, from providing habitat for terrestrial and aquatic organisms, playing a role in nutrient cycling and carbon cycles, and affecting stream flows. By studying the litter and woody debris in the Niwot Ridge LTER station, we may be able to infer what types of organisms and relationships exist in that area.

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.Samples are taken at terrestrial sites with woody vegetation >2 m tall, at randomly selected tower plots. 2.Frequency of plot sampling depends on the type of site. Gound traps are sampled once per year, while deciduous sites are sampled more frequently than evergreen sites. 3. The finest resolution of data is one trap. This contains a unique trapID. It should also tell you the subplot, plot, site, and domain.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
#Find dimensions of Neonics
dim(Neonics)
```

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

```r
#the dimensions are 30 columns, 4623 rows
```

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? [Tip: The `sort()` command is useful for listing the values in order of magnitude...]

```r
sort(summary(Neonics$Effect))
```

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

Answer:It appears the top two most common effects studied were population and mortality. These effects might specifically be of interest if researchers want to understand how these pesticides affect insect populations overall and what effects they have on insect mortality to inform best practices around applying this class of insecticides.

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: Explore the help on the `summary()` function, in particular the `maxsum` argument...]

```r
sort(summary(Neonics$Species.Common.Name, maxsum=7))
```

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

Answer:The six most common species were honey bee, parasitic wasp, buff tailed bumblebee, carniolan honey bee, bumble bee, and italian honeybee (excluding "other"). These are all pollinators.

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? [Tip: Viewing the dataframe may be helpful...]

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

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

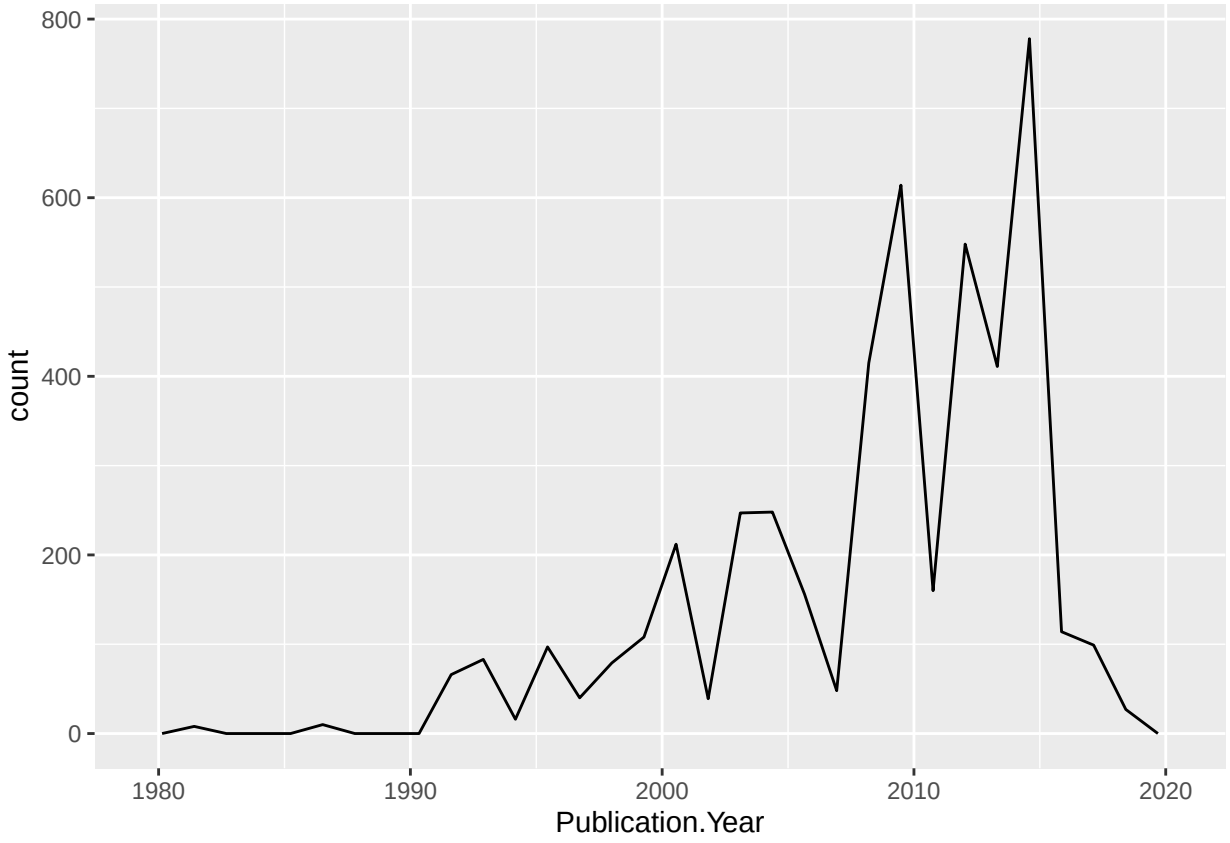
Answer:"Conc.1...Author" is a factor that shows the value of the concentration. It is a factor and not numberic because some of the values have backslash and others have NR, so R recognizes it as a factor.

Explore your data graphically (Neonics)

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

```
#Plot Number of Studies
ggplot(Neonics)+
  geom_freqpoly(aes(x=Publication.Year))
```

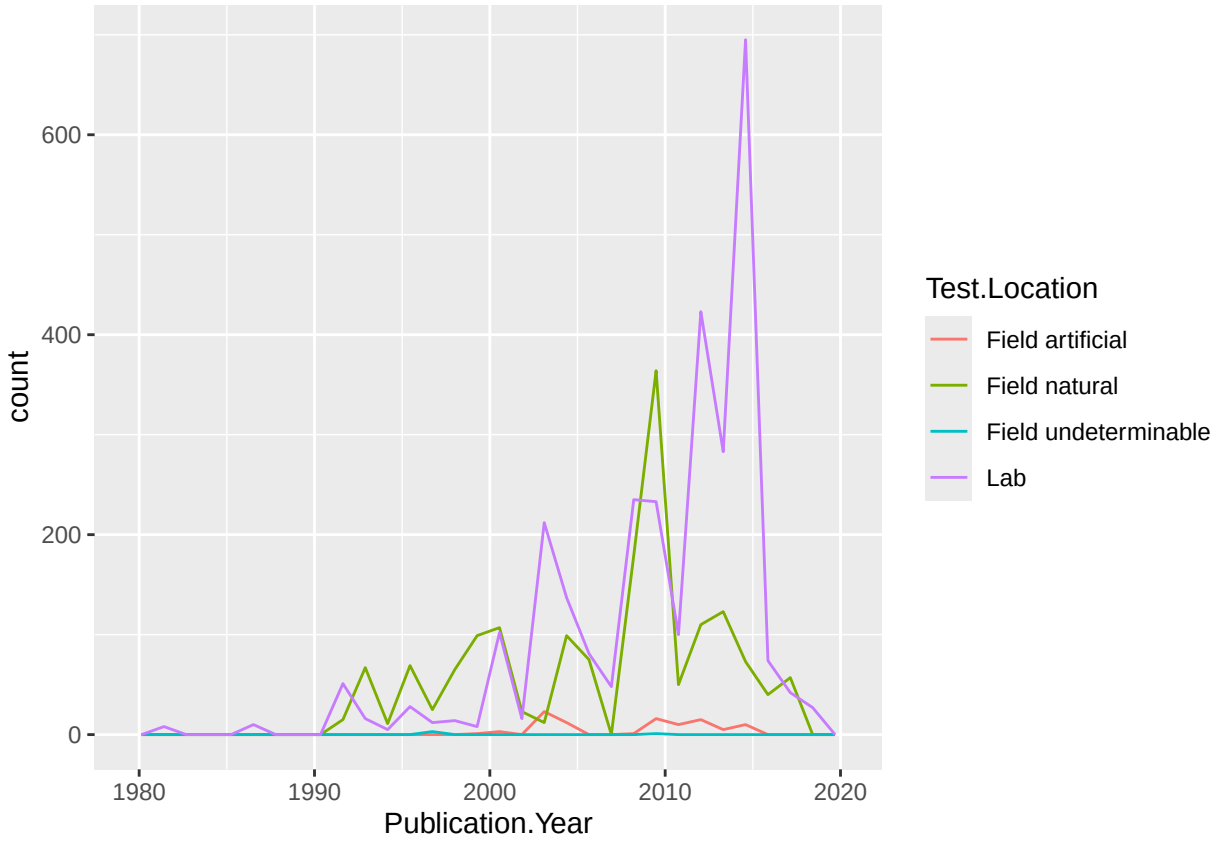
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
#Add Color to Graph
ggplot(Neonics)+
  geom_freqpoly(aes(x=Publication.Year, color=Test.Location))
```

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



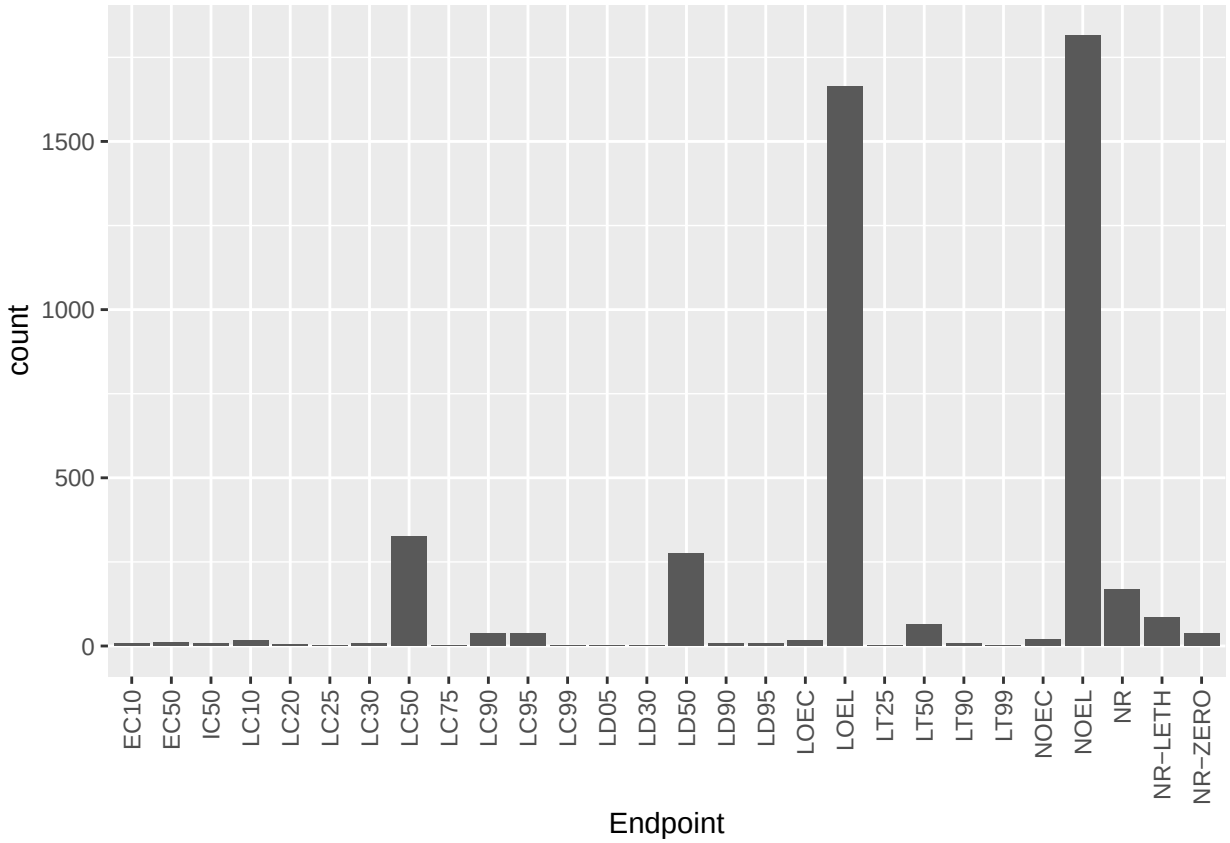
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer:The most common test locations were lab and field natural. Over time, it seems lab test locations became more frequent, especially after 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...]

```
#Create bar graph of endpoints
ggplot(Neonics)+
  geom_bar(aes(x=Endpoint))+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer:The two most common are LOEL and NOEL. LOEL = Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different from controls. NOEL = No-observable-effect-level: highest dose (concentration)producing effects not significantly different from 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"
```

```
#convert to a date
Litter$collectDate <- ymd(Litter$collectDate)
class(Litter$collectDate)
```

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

```
unique(Litter$collectDate, incomparables = FALSE)
```

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

```
#litter was sampled on 2018-08-02 and 2018-08-30
```

13. Using the `unique` function, determine how many different plots were sampled at Niwot Ridge. How is the information obtained from `unique` different from that obtained from `summary`?

```
#Determine number of unique plots sampled
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
```

```
# Unique shows 12 levels
summary(Litter$plotID)
```

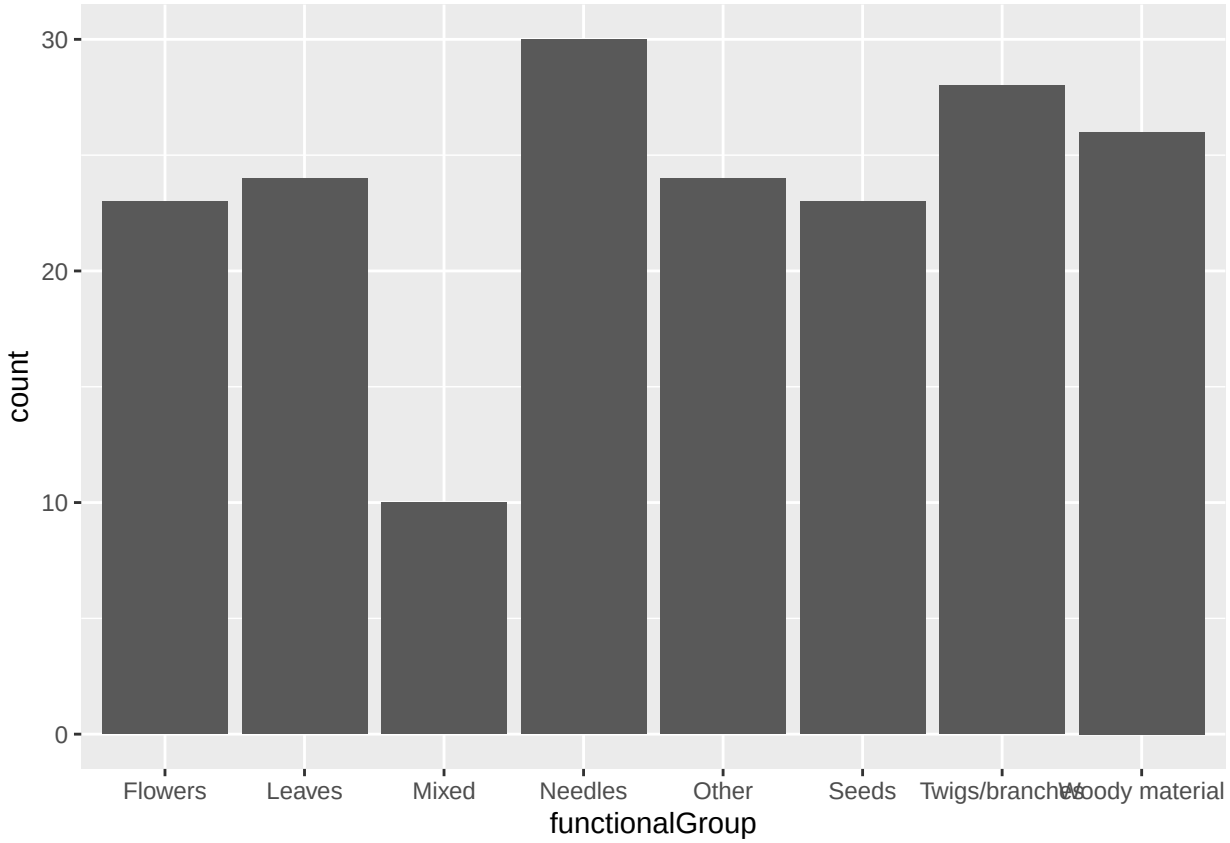
```
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061
##       20       19       18       15       14        8       16       17
## NIWO_062 NIWO_063 NIWO_064 NIWO_067
##       14       14       16       17
```

```
#Summary shows the count of samples at each of the 12 plots
```

Answer:Summary shows the count of samples at each of the plots, whereas unique shows just the levels.

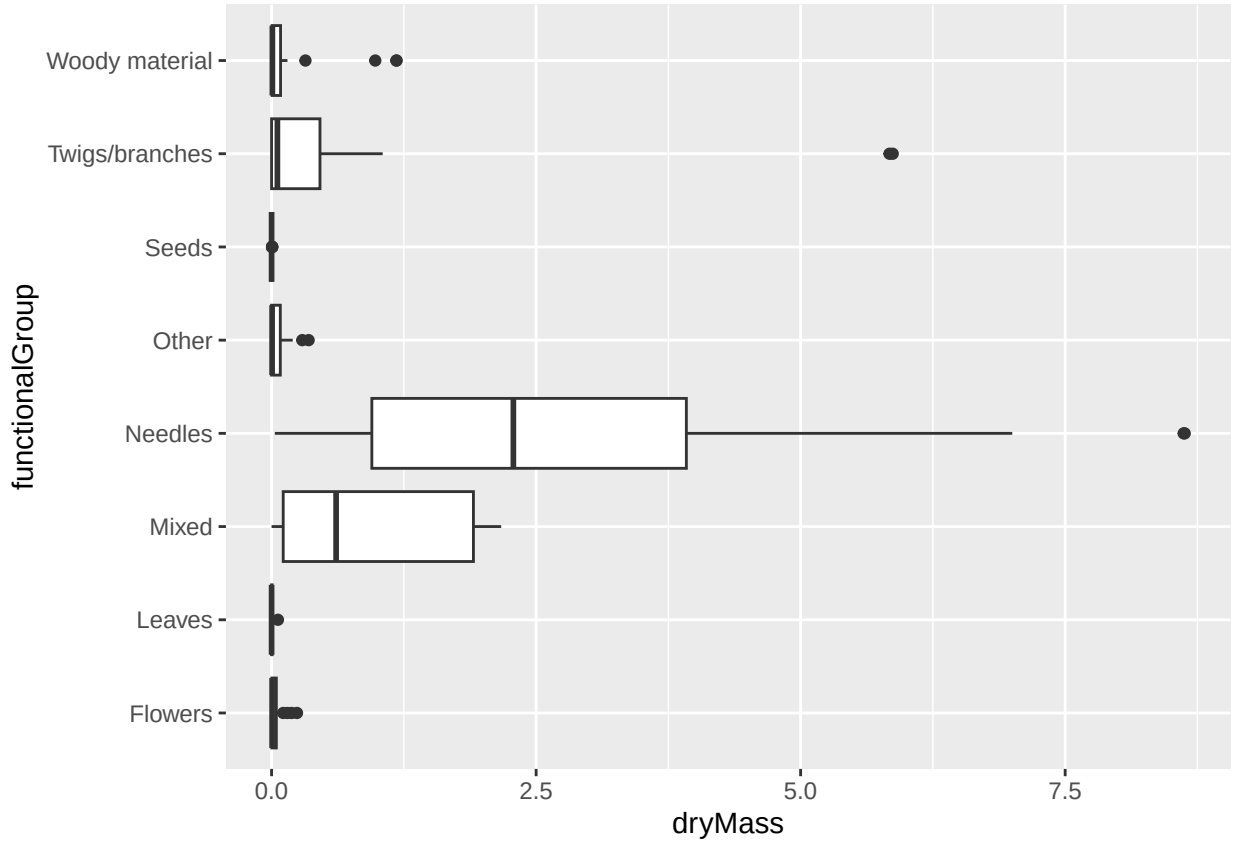
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#Create bar graph of functional group counts
ggplot(Litter)+
  geom_bar(aes(x=functionalGroup))
```

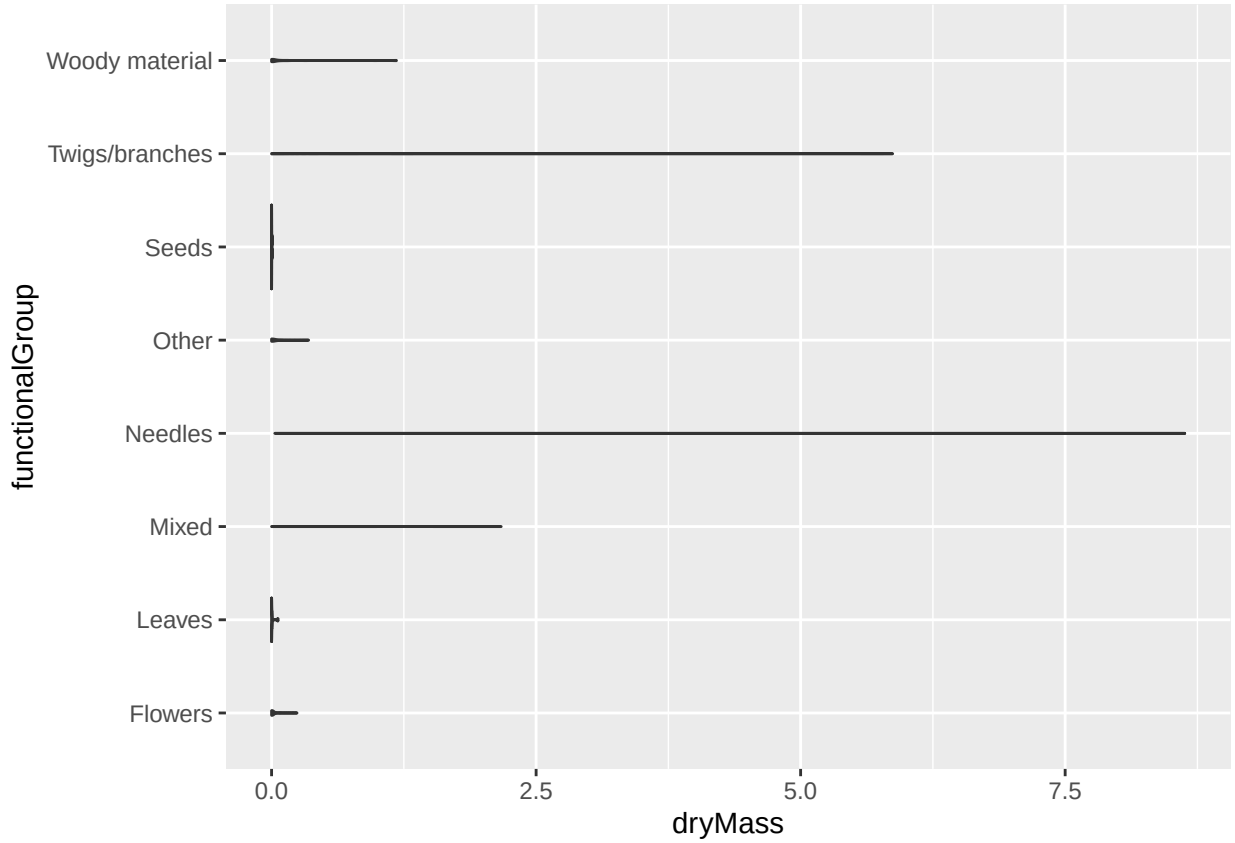


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#Create boxplot and violin plot
ggplot(Litter)+
  geom_boxplot(aes(x=dryMass, y=functionalGroup))
```




```
ggplot(Litter)+
  geom_violin(aes(x=dryMass,y=functionalGroup))
```



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

> Answer:The violin plot displays density distributions, but it is not effective because there are not enough datapoints at each unique value of dry mass to show density. The boxplot is better at showing distribution.

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

> Answer:Needles