Assignment 3: Data Exploration

Vicky Fong

Sep 23, 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.

```r
#Load packages
library(tidyverse)
library(lubridate)
library(here)

#Check working directory
getwd()
```

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

```r
here()
```

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

```r
#Upload datasets
neonics <- read.csv(
  here('Data','Raw','ECOTOX_Neonicotinoids_Insects_raw.csv'), stringsAsFactors = T
)

litter <- read.csv(
  here('Data','Raw','NEON_NIWO_Litter_massdata_2018-08_raw.csv'), stringsAsFactors = T
)
```

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: Research has shown that neonicotinoid insecticides have detrimental ecological impcats, including being very toxic to pollinators, beneficial insects, and aquatic invertebrates. Pollinators are especially important ecologically as they are crucial in the reproduction of many plants.

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 are important part of terrestrial and freshwater ecosytems, including: contribute to carbon budgets and nutrient cycling, source of energy for aquatic ecosystems, habitat for terrestriaal and aquatic organisms, influence water flow and sediment transport.

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. Litterfall is collected from elevated traps and fine woody debris is collected from ground traps. 2. Mass is measured separately for each of the functional groups to an accuracy of 0.01 grams. 3. Sampling is only done at terrestrial NEON sites that contain woody vegetation >2m tall in tower plots.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
nrow(neonics)
```

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

```
ncol(neonics)
```

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

```
#The neonics dataset has 4623 observations of 30 variables.
```

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...]

```
#Sort effect column by decreasing values
sort(summary(neonics$Effect),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 that are studied are population and mortality. We want to understand how insecticides influences population sizes and the mortality of different insect species.

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...]

```
#Sort species common name column by decreasing values with maximum of 7 categories (6 + other)
sort(summary(neonics$Species.Common.Name, maxsum = 7),decreasing = TRUE)
```

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

> Answer: The six most commonly studied species are Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. All of these species are pollinators that serve important ecological roles in plant reproduction and thus overall health of an ecosystem.

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.)
```

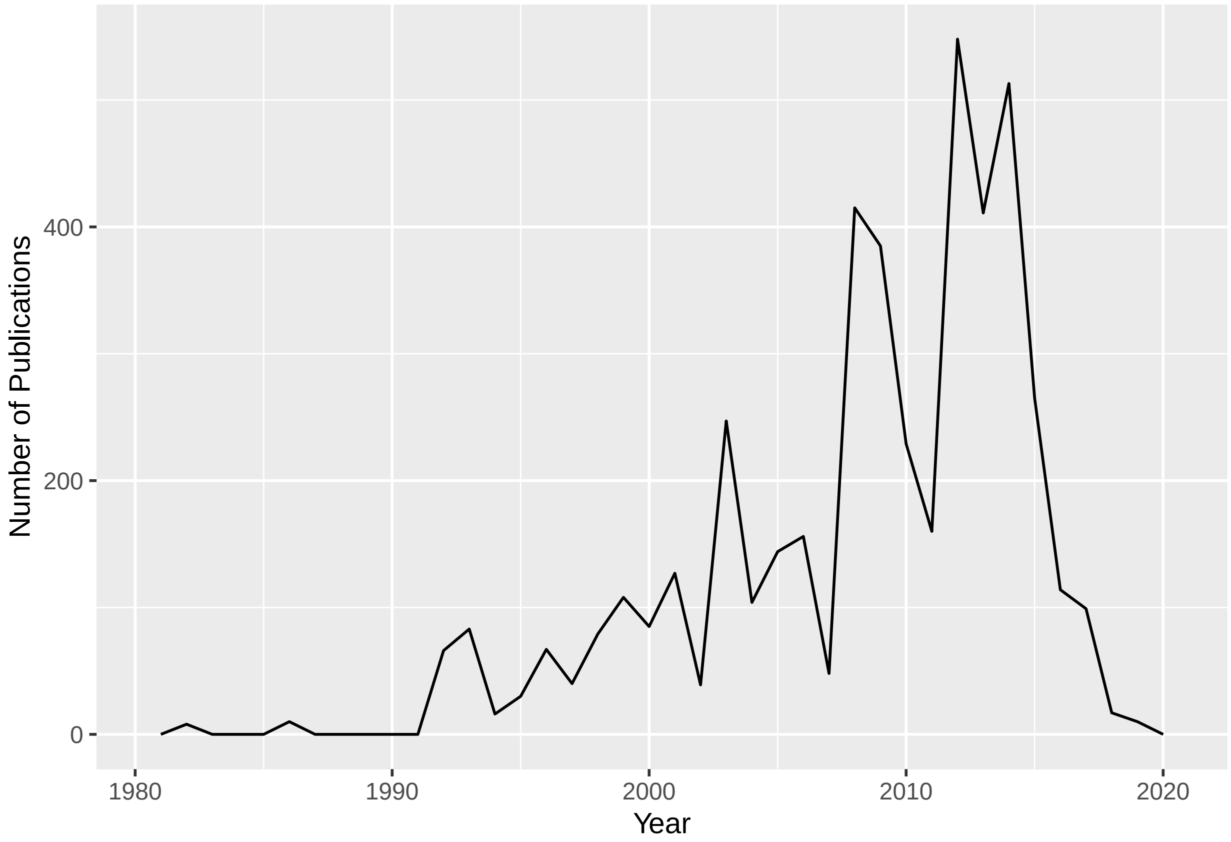
```
## [1] "factor"
```

Answer: This column is not numeric as it contains rows where concentrations where not recorded ("NR") or estimated ("~10").

Explore your data graphically (Neonics)

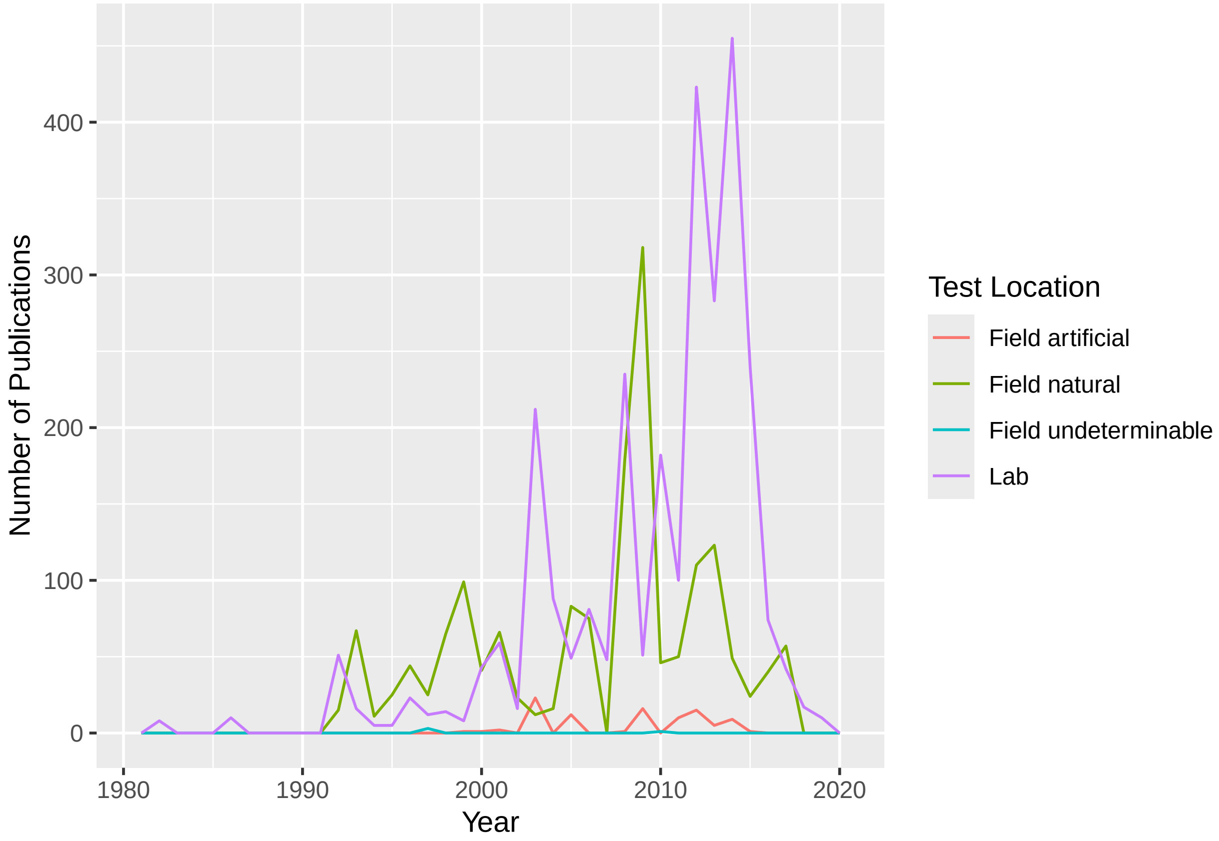
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
#Plot of publications by year
ggplot(neonics,aes(x=Publication.Year)) +
  geom_freqpoly(binwidth=1) +
  labs(x='Year', y='Number of Publications',color='Test Location')
```



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

```r
#Plot of publications by year and test location
ggplot(neonics,aes(x=Publication.Year, color=Test.Location)) +
  geom_freqpoly(binwidth=1) +
  labs(x='Year', y='Number of Publications',color='Test Location')
```



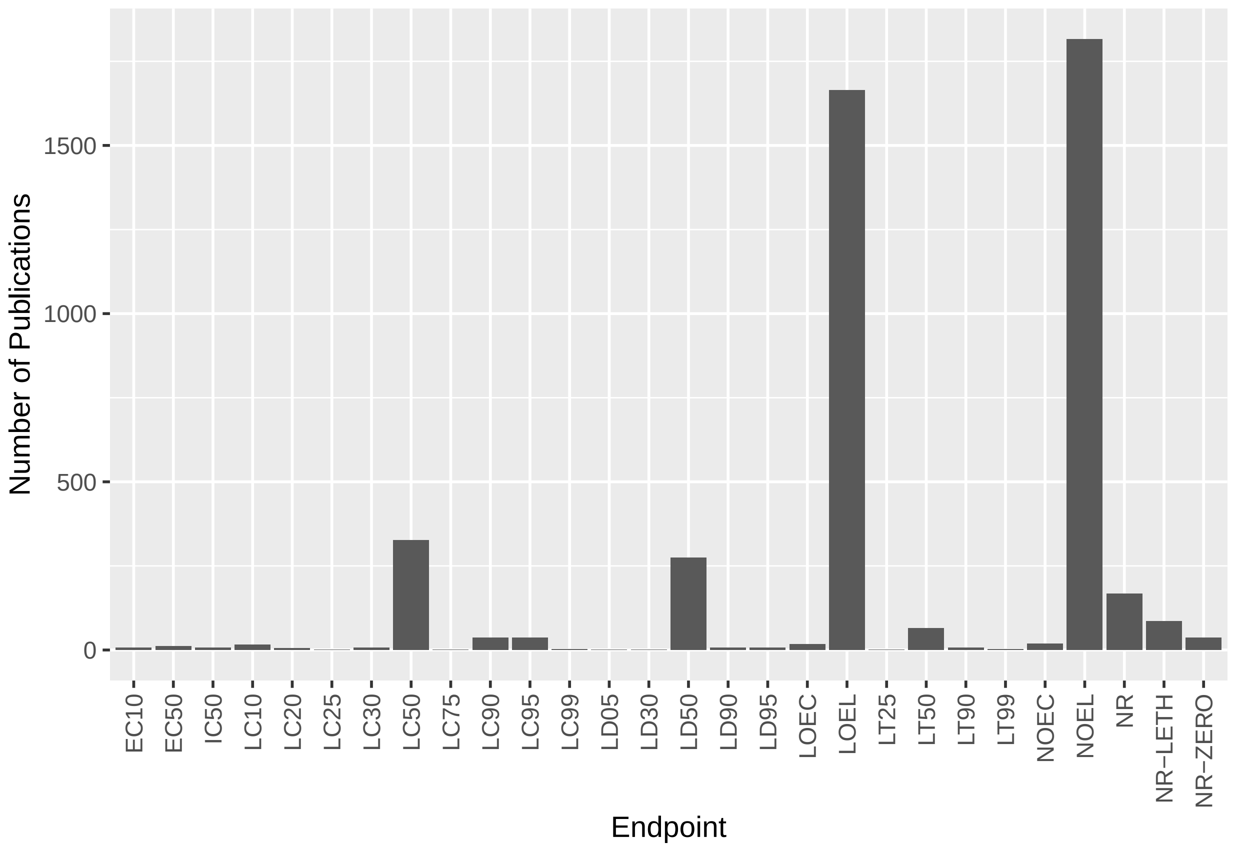
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are in the lab between 2010 and 2020. The next most common test location is natural field between 2005 ans 2015.

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…]

```r
#Bar graph of Endpoint counts
ggplot(neonics,aes(x=Endpoint)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  labs(x='Endpoint',y='Number of Publications')
```



Answer: The two most common endpoints are NOEL (No-Observable-Effect-Level: highest dose producing effects not siginificantly different from control) and LOEL (Lowest-Observable-Effect-Level: lowest dose producing effects were significantly different from control).

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.

```
#Change class of collectDate column
class(litter$collectDate)
```

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

```
litter$collectDate <- ymd(litter$collectDate)
class(litter$collectDate)
```

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

```
#Extract day
litter$day <- day(litter$collectDate)

#Which dates in August 2018
unique(litter$day)
```

```
## [1]  2 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`?

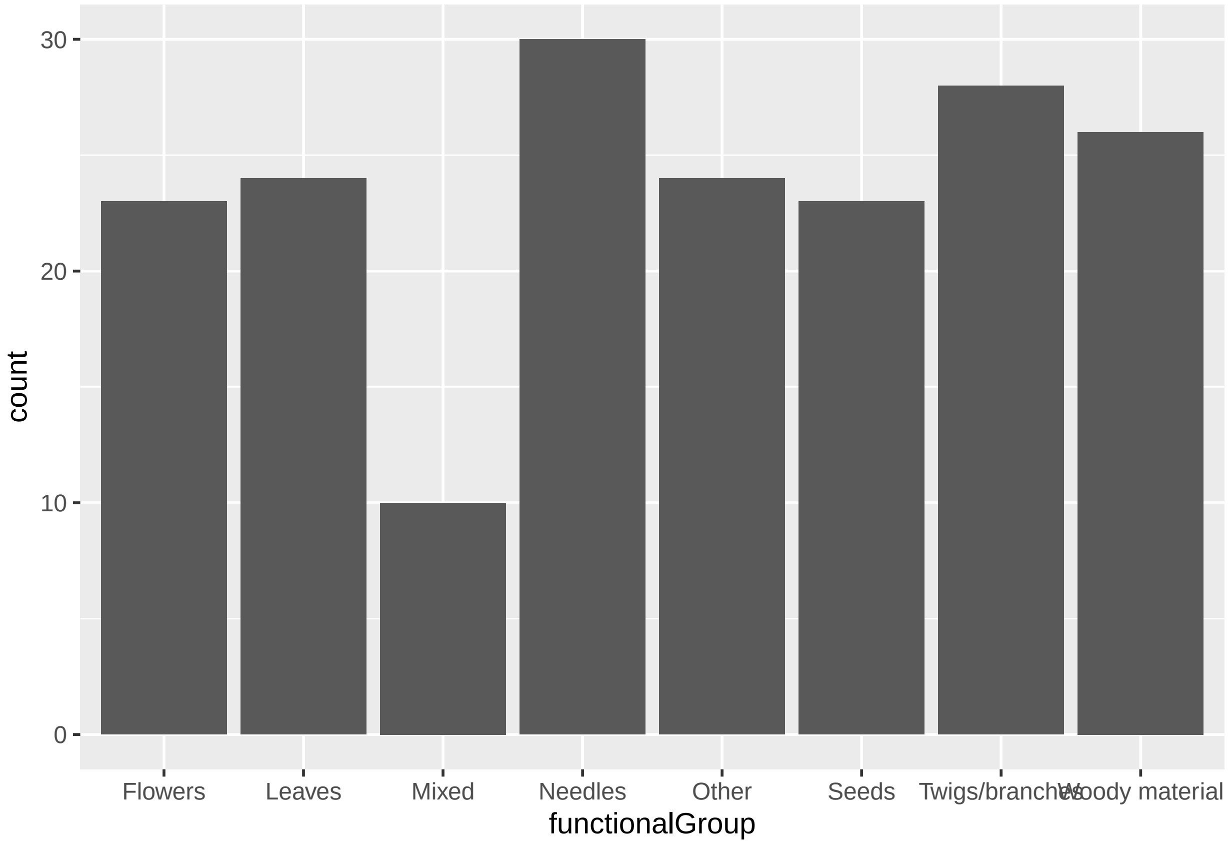
```
unique(litter$plotID)
```

```
##  [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: unique() does not return the count of each plotID and only tells you the number of unique values there are.

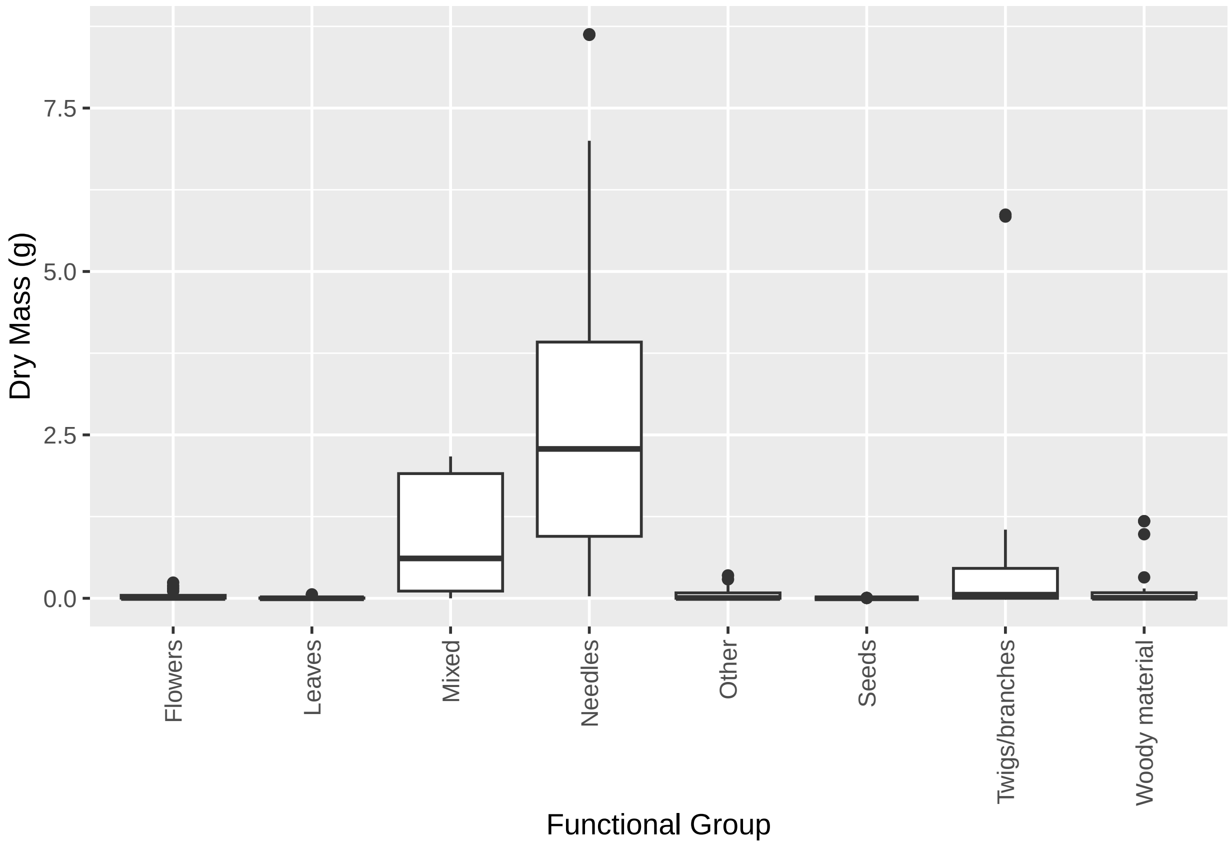
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(litter,aes(x=functionalGroup)) +
  geom_bar()
```

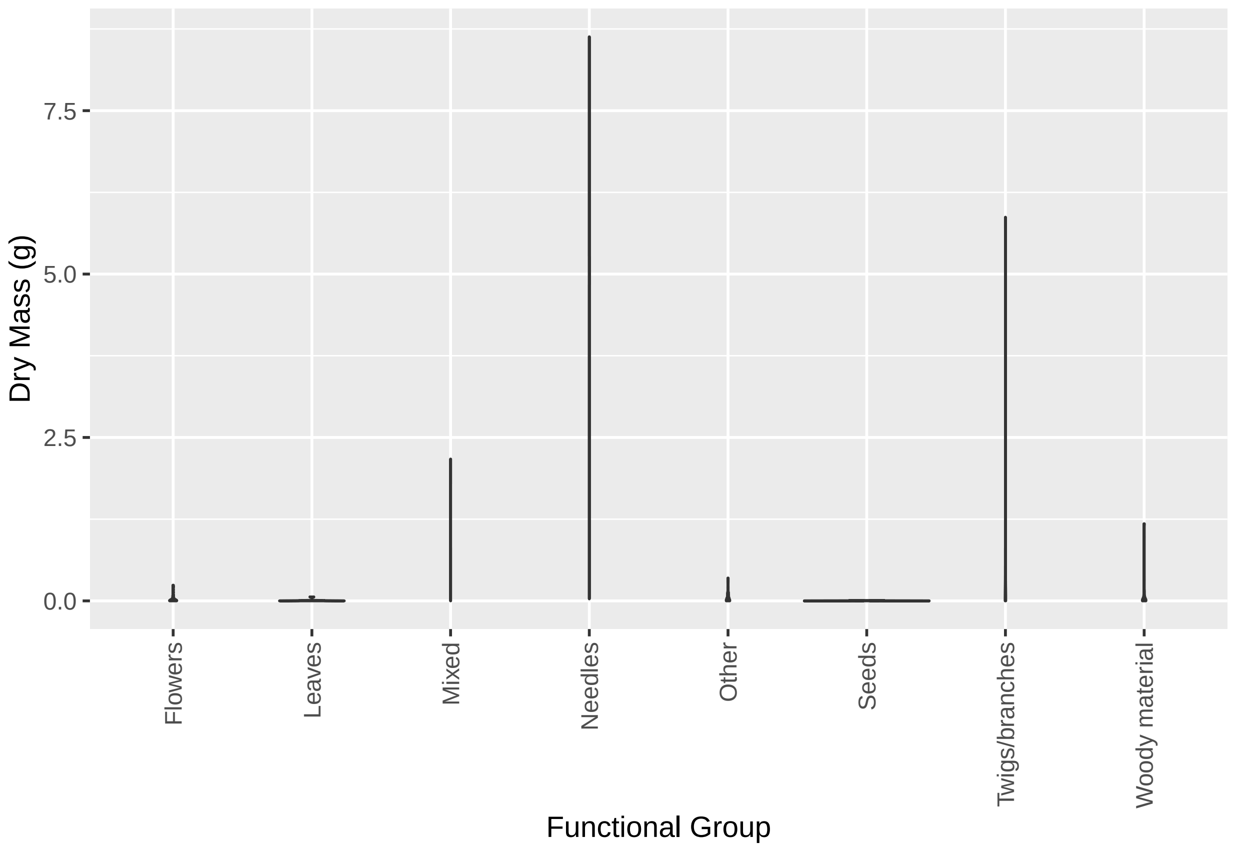


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(litter,aes(x=functionalGroup, y=dryMass)) +
  geom_boxplot() +
  labs(x='Functional Group', y='Dry Mass (g)') +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



```
ggplot(litter,aes(x=functionalGroup, y=dryMass)) +
  geom_violin() +
  labs(x='Functional Group', y='Dry Mass (g)') +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



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

> Answer: Since there 8 functional groups, there is not enough space on the x-axis to fully show the details of each violin plot, but the boxplot is able to show the differences in range and average for each functional group.

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

> Answer: Needles have the highest biomass at these sites, followed by mixed litter.