Assignment 3: Data Exploration

Lucy Wang

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.

```
# check working directory
getwd()
```

```
## [1] "/Users/lucywang/Documents/EDE_Fall2023"
```

```
#load packages
library(tidyverse)
library(lubridate)

# read datasets as dataframes, read strings as factors
Neonics <- read.csv('./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv', stringsAsFactors = T)
Litter <- read.csv('./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:The use of neonicotinoids may discrupt the ecosystem balance and harm the biodiversity. Because many insects including important pollinators depend on farmlands as a part of their ecosystem, and other species (including human) depend on the vegetation pollinated by those insects. If neonicotinoids are killing insects that have positive impacts, the ecosystem may face a risk of collaspe. Therefore, we are interesed in studying the ecotocicology of neonicotinoids to prevent that from happening.

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 plays an important role in the carbon cycle or the nutrient cycle for soil and plantations. The litterfall breaks down into soil organic matter and provide nutrients for plants. Therefore, studing litter and woody debris can help learning the forest biochemical cycle and the health of forests in general.

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. The sampling is done at NEON sites, and the litter is divided into 8 function groups. 2. The spatial relationship between traps, subplots, plots, sites, and domains explains the coding method for locations. 3. The sampling is done at NEON sites taht have woody vegetations above 2 meters tall. There are forested sites and sites with low-statured vegetation.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
colnames(Neonics) #call out the column names of the dataset
```

```
##  [1] "CAS.Number"                    "Chemical.Name"
##  [3] "Chemical.Grade"                "Chemical.Analysis.Method"
##  [5] "Chemical.Purity"               "Species.Scientific.Name"
##  [7] "Species.Common.Name"           "Species.Group"
##  [9] "Organism.Lifestage"            "Organism.Age"
## [11] "Organism.Age.Units"            "Exposure.Type"
## [13] "Media.Type"                    "Test.Location"
## [15] "Number.of.Doses"               "Conc.1.Type..Author."
## [17] "Conc.1..Author."               "Conc.1.Units..Author."
## [19] "Effect"                        "Effect.Measurement"
```

```
## [21] "Endpoint"                          "Response.Site"
## [23] "Observed.Duration..Days."          "Observed.Duration.Units..Days."
## [25] "Author"                            "Reference.Number"
## [27] "Title"                             "Source"
## [29] "Publication.Year"                  "Summary.of.Additional.Parameters"
```

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) # Summarize 'Effect' column
```

```
##     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: The most common effects that are studied are Poplulation (1803) and Mortality (1493). Neonicotinoids may lead to vital effects that impose significant threats on the eco-health and food security for huamns. Decrease in insect population and insects' deaths are the two most direct causes to ecological harm and uncentain food production for all 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: The `sort()` command can sort the output of the summary command...]

```
# Summarize species common name in the dataset. Assign to a dataframe called 'Species.data'
Species.data <- as.data.frame(summary(Neonics$`Species.Common.Name`))

# list 6 most common studied species in the dataset
head(Species.data,6)
```

```
##                       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: The six most commonly studies species in the dataset are Honey Bee (667), Parasitic Wasp (285), Buff Tailed Bumblebee (183), Carnoiolan Honey Bee (152), Bumble Bee (140), and Italian Honeybee (113). The commonality among them is that they are all bees. Because bees are the most common pollinators that are essential for dispersing plant seeds. Without pollinators, the crop yield would shrink significantly. In addition, neonicotinoids can impose major health impacts on bees due to the toxic chemicals.

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?

```
# check the class of 'Conc.1..Author'
class(Neonics$Conc.1..Author)
```

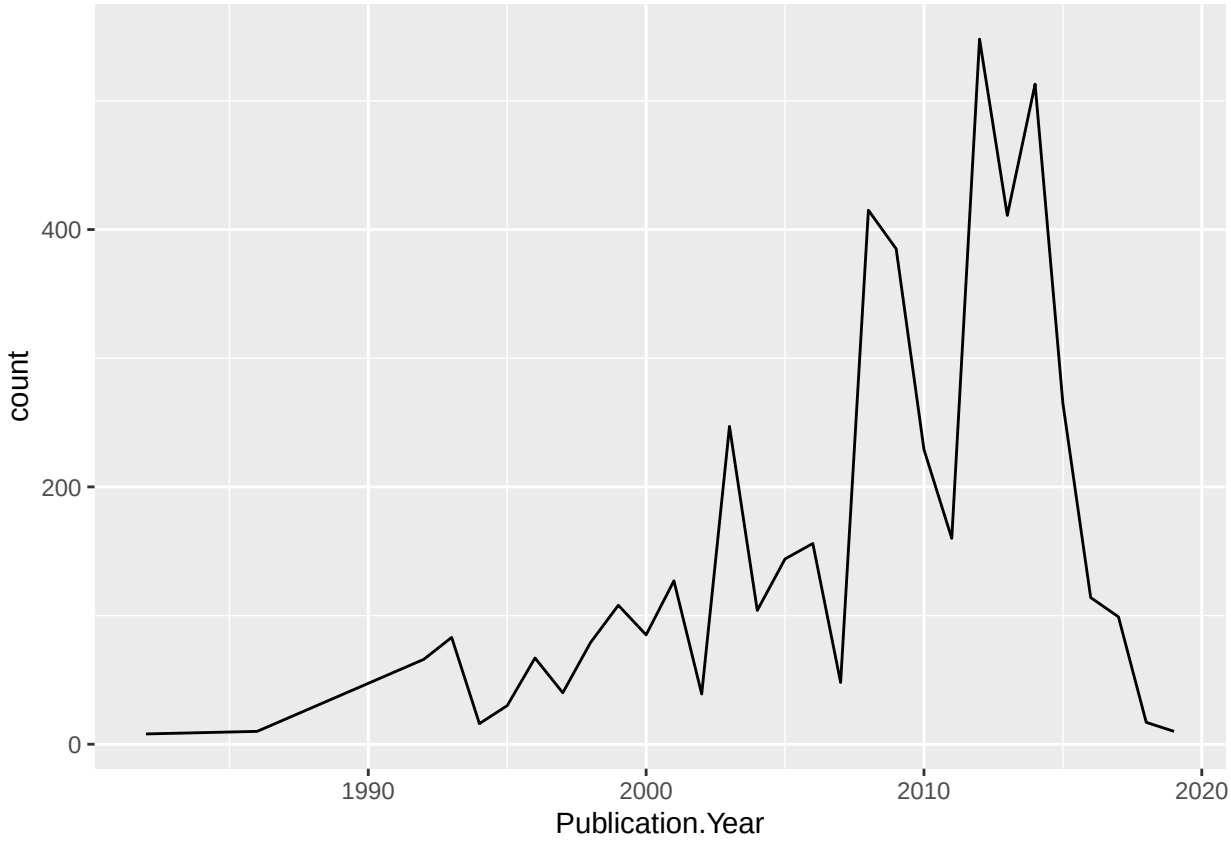
```
## [1] "factor"
```

Answer: The class of `Conc.1..Author.` is Factor. It is not numeric because it includes certain values like "NR", "144/", "~10" that contains non-numeric symbols and characters.

Explore your data graphically (Neonics)

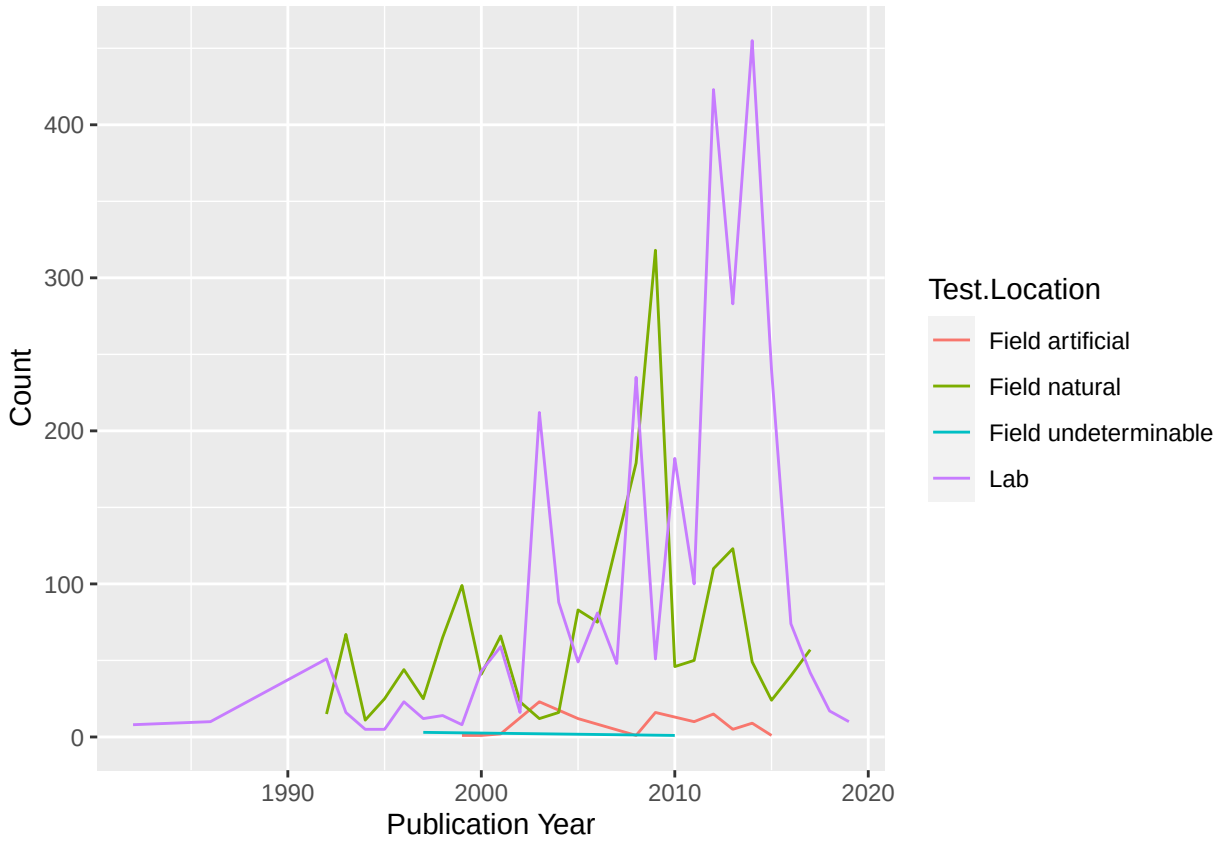
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
# generate a frequency polygon that shows the number of studies conducted by publication year
ggplot(Neonics) +
  geom_freqpoly(aes(x=Publication.Year),stat='count')
```



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

```
# generate a frequency polygon that shows the number of studies conducted by publication year, dispayed
ggplot(Neonics) +
  geom_freqpoly(aes(x=Publication.Year, color = Test.Location),stat='count') +
  xlab('Publication Year')+
  ylab('Count')
```



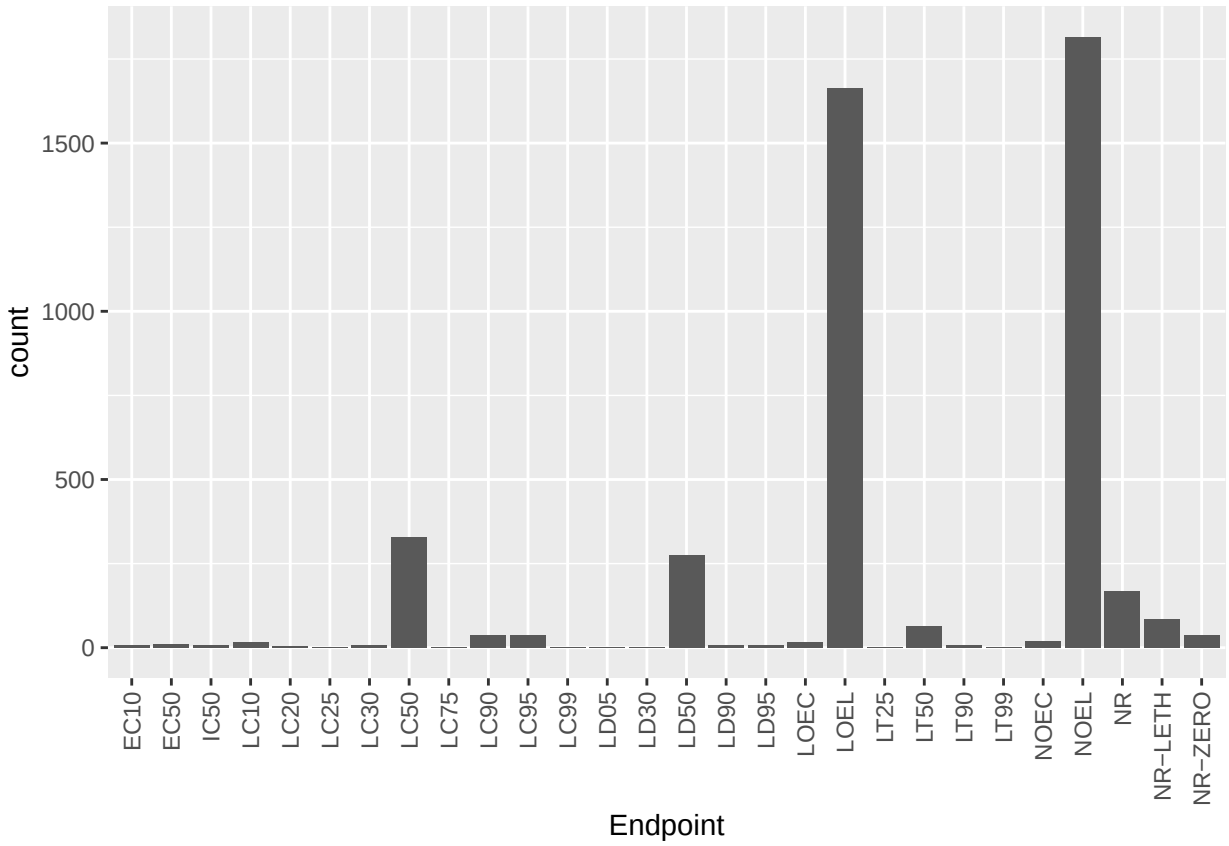
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test location is Lab, followed by Field natural. Lab has the longest history and is the most common during the early 2000s and peaked during 2010-2015. Other loctaions did not come up until the 1990s. Field natural became the the most common test location until it was taken over by Lab in the early 2000s. However, it continued to grow and peaked in late 2000s before the rapid growth of Lab. Field underterminable and Field artificial remained relatively uncommon throughout the timespan and have not been in use in recent years.

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...]

```
# generate a bar graph of Enpoint counts, adjust the x-axis lables to fit the graph.
ggplot(Neonics) +
  geom_bar(aes(x=Endpoint), stat='count') +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



Answer: The most common end points are NOEL and LOEL, which is each defined as no-observable-effect-level and lowest-observable-effect-level for terrestrial database usage.

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.

```
# the class of collectDate was a factor. Change to Date.
Litter$collectDate <- ymd(Litter$collectDate)
class(Litter$collectDate) # test the class again
```

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

```
# Find which dates litter was sampled in 2018/08
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`?

```r
# Find how many plots were sampled at Niwot Ridge
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
```

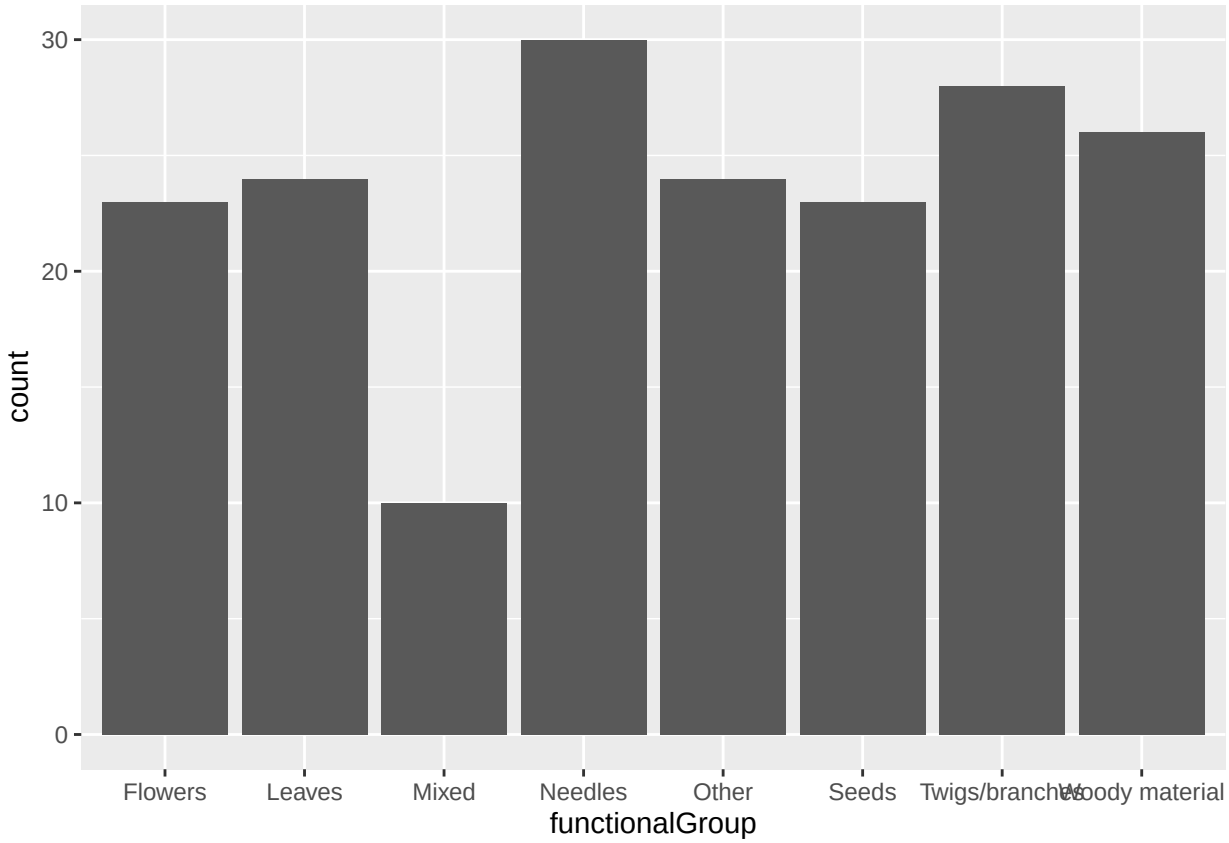
```r
summary(Litter$plotID)
```

```
## 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
```

Answer: 12 plots were sampled at Niwot Ridge. The 'unique' function shows the "names" of plotIDs and the amount of 'levels' (which, in this case, is the number of plots), while the 'summary' function shows the count of smaples at each plot insead of showing the number of plots.

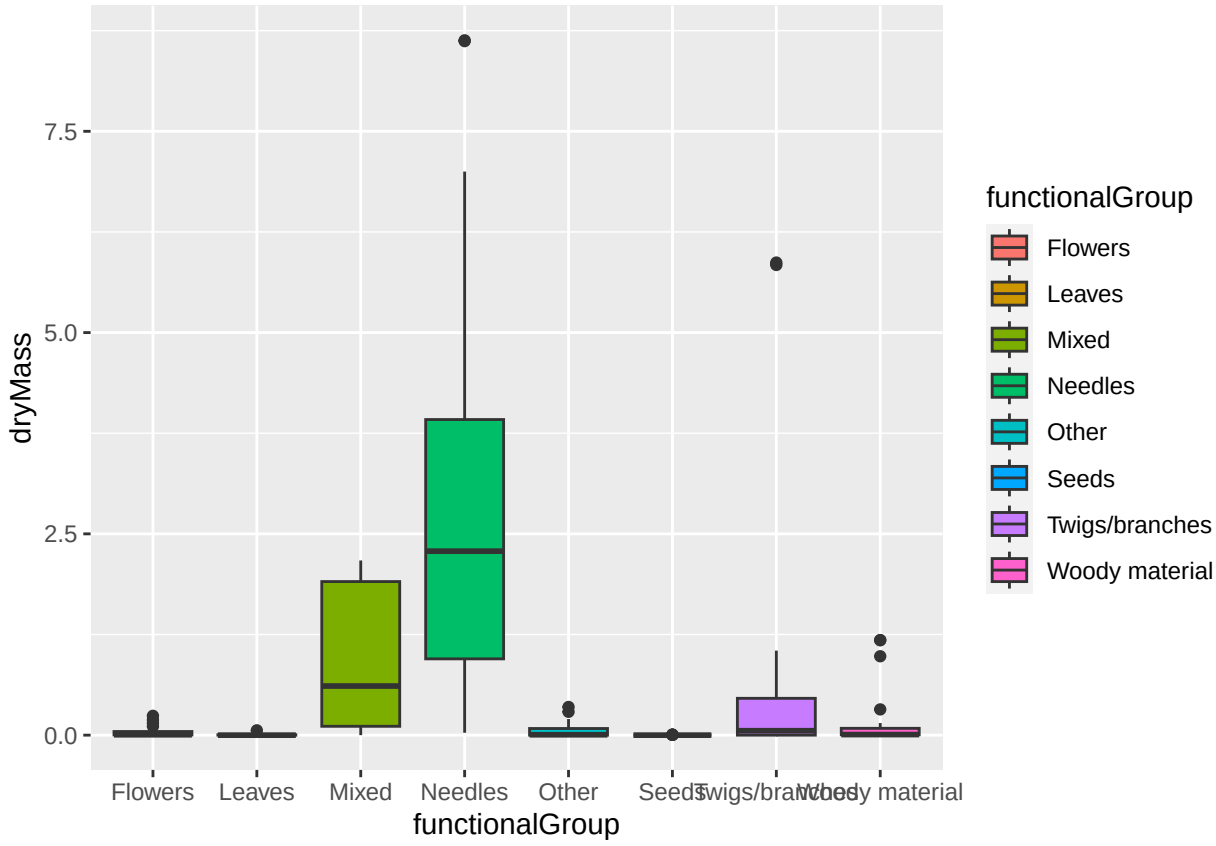
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
# generate a bar graph of functionalGroup counts to show the types of litter collected
ggplot(Litter)+
  geom_bar(aes(x=functionalGroup), stat = 'count')
```

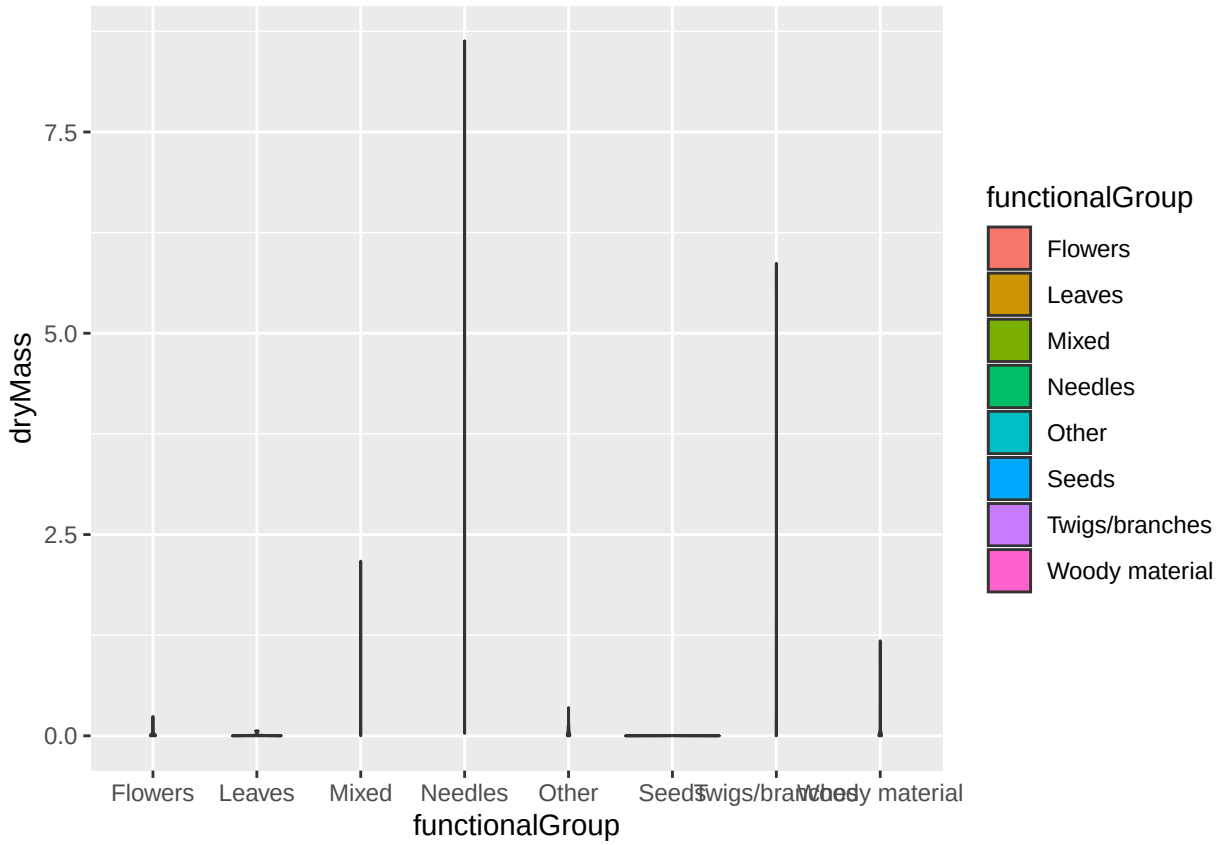


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
# generate a boxplot of dryMass by functionalGroup
ggplot(Litter)+
  geom_boxplot(aes(x=functionalGroup, y=dryMass, fill = functionalGroup))
```




```
# generate a violin graph of dryMass by functionalGroup
ggplot(Litter)+
  geom_violin(aes(x=functionalGroup, y=dryMass, fill = functionalGroup))
```



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

> Answer: Because the bymass data is nearly distinct for each sample, and data range so wide for different functional groups but some functional groups that have much smaller mass than others. The violin plot shows the distribution density of data, so it is not suitable for such a scattered dataset that contains disctinct numbers and extreme, varied ranges. The boxplot, on the other hand, only reflects the interquartile range without reflecting the distribution curve. So the scatteredness and distinction of numbers do not affect boxplots.

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

> Answer: Needles tend to have the highest biomass, followed by mixed litter and branches, due to the ranking in dryMass.