Assignment 3: Data Exploration

Jack Eynon

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Exploration.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Salk_A03_DataExploration.Rmd") prior to submission.

The completed exercise is due on Tuesday, January 28 at 1:00 pm.

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.

```
library(tidyverse)
Neonics <- read.csv("~/Courses/EnvDataAnalytics/Environmental_Data_Analytics_2020/Data/Raw/ECOTOX_Neoni
Litter <- read.csv("~/Courses/EnvDataAnalytics/Environmental_Data_Analytics_2020/Data/Raw/NEON_NIWO_Lit
```

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 ecotoxicologoy of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: Neonicotinoids might be detrimental to ecosystem health, particularly if insects are an important source of energy for higher trophic species (i.e. fish, mammals) or serve other important functions (e.g. pollination) in that ecosystem. There might also be human health risks if neonicotinoids bioaccumulate in species we eat.

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: We might be interested in studying forest litter and woody debris in order to learn more about fuel loading and wildfire risk in Colorado. Or perhaps to learn about habitat suitability for particular species.

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: * Litter and woody debris is collected from elevated and ground traps and weighed to within 0.01 grams. * Locations of the sampling plots are selected randomly within the 90% flux footprint of the primary and secondary airsheds, and trap placements within plots may be either targeted or randomized, depending on the vegetation. * Ground traps are sampled once per year, and elevated traps are sampled more frequently.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

6. Using the `summary` function, 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: The most common effects that are studied are population and mortality. These effects might be of particular interest if researchers are interested in studying whether neonicotinoid concentrations reduce insect abundance, induce trophic cascades, etc.

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.

```
head(summary(Neonics$Species.Common.Name))
```

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

Answer: The 6 most studied species are all bees or wasps. They might be of interest over other insects because of their important ecological roles. Wasps are predatory species that keep the populations of other insects in check (including some that are pests for agriculture). As pollinators, bees support wild plant reproduction and growth.

8. Concentrations are always a numeric value. What is the class of Conc.1..Author. in the dataset, and why is it not numeric?

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

```
##  Factor w/ 1006 levels "<0.0004","<0.025",..: 639 510 813 622 442 637 500 642 814 784 ...
```

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

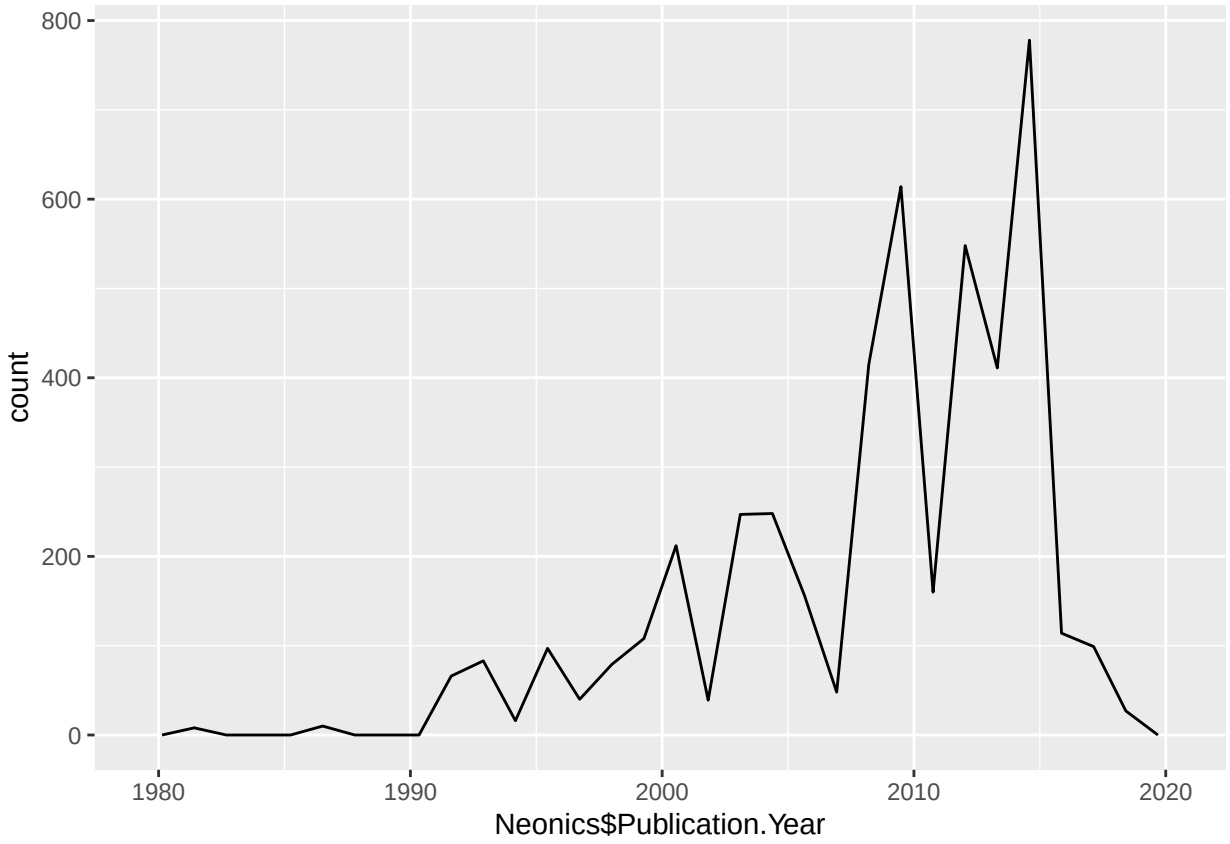
```
## [1] "factor"
```

Answer: The data is in a factor (categorical) class. It is not numeric because the units are not standardized across the values in that column; therefore, numeric summaries (e.g. descriptive statistics) of the data would not be meaningful. Keeping the data as factor values rather than numeric values might be useful in preventing people from misinterpreting the data.

Explore your data graphically (Neonics)

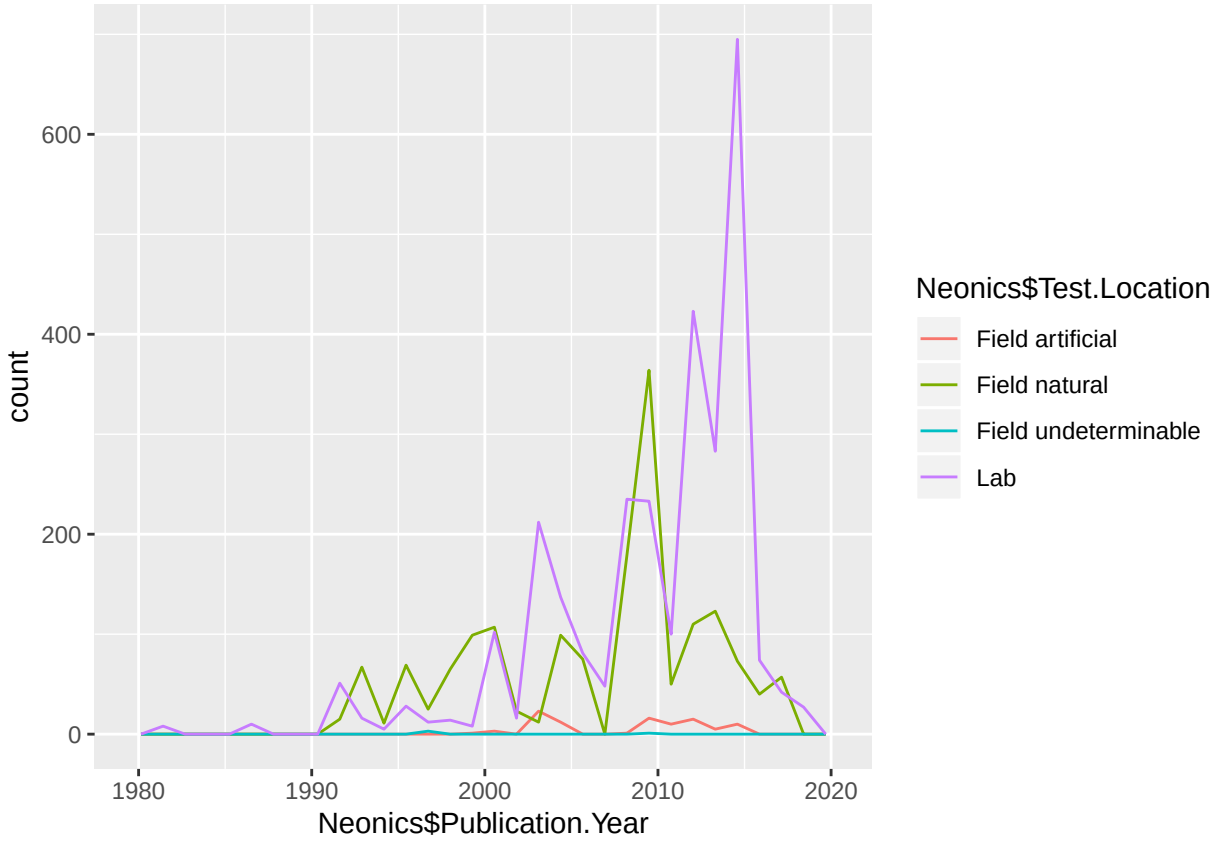
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(data = Neonics, aes(x=Neonics$Publication.Year)) +
  geom_freqpoly(bins=30)
```



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

```
ggplot(data = Neonics, aes(x=Neonics$Publication.Year, color = Neonics$Test.Location)) +
  geom_freqpoly(bins=30)
```

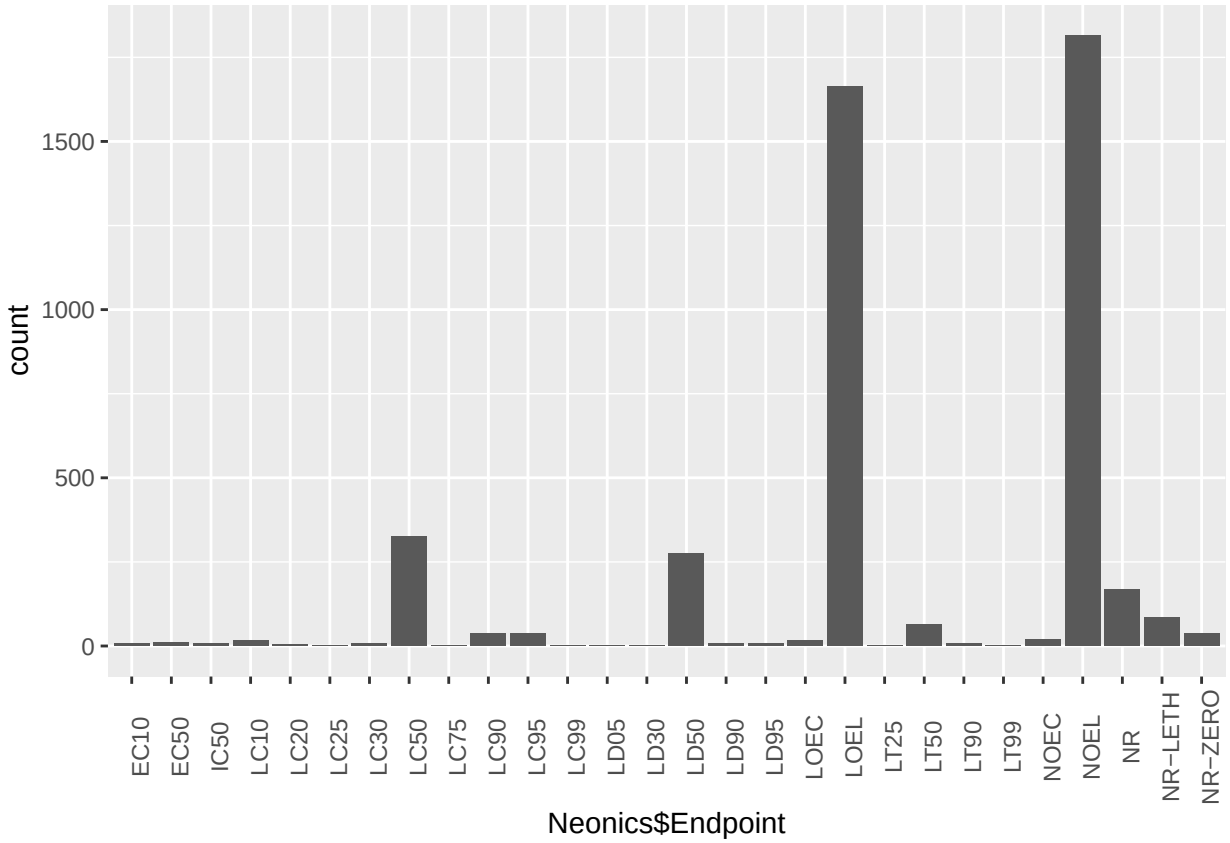


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are "Lab" and "Field Natural". The number of studies with a "Field Natural" location peaked just before 2010, when it was the most popular study location. The number of studies done in a "Lab" setting peaked in the 2010s, and over that period "Lab" was by far the most common study location.

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(data = Neonics, aes(x=Neonics$Endpoint)) +
  geom_bar() + theme(axis.text.x = element_text(angle = 90))
```



Answer: The two most common endpoints are NOEL (No-observable-effect-level), which is defined as the highest dose producing effects not significantly different from responses of controls, and LOEL (Lowest-observable-effect-level), which is defined as the lowest dose producing effects that were significantly different from responses of 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) # Factor
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate)
unique(Litter$collectDate) # August 2nd and August 30th
```

```
## [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`?

```
summary(Litter$siteID) #188
```

```
## NIWO
##  188
```

```
unique(Litter$siteID)
```

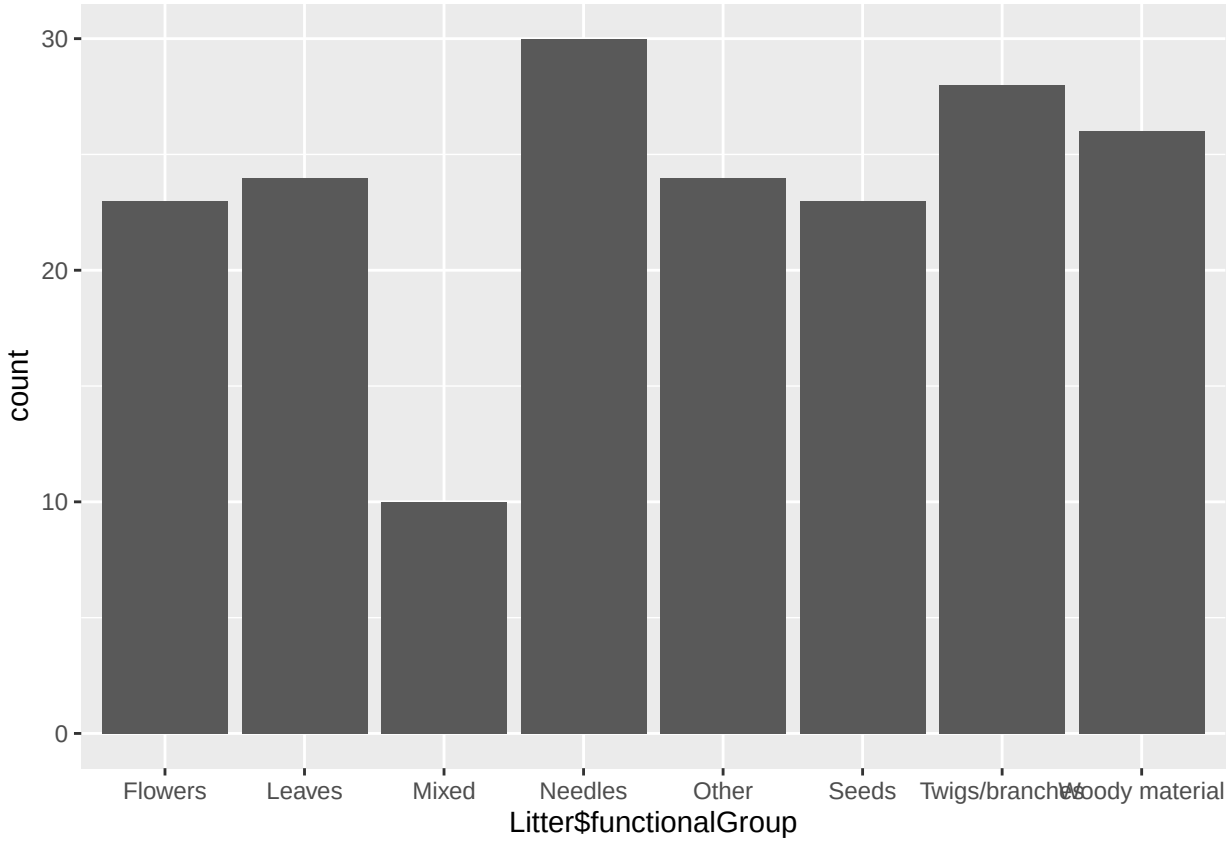
```
## [1] NIWO
```

```
## Levels: NIWO
```

Answer: The 'unique' function lists all discrete values of the data that is referenced. Using the 'unique' function on the siteID column, we know that there is only one unique value for siteID, which is NIWO. Therefore, we know that all 188 observations were sampled at NIWO (Niwot Ridge).

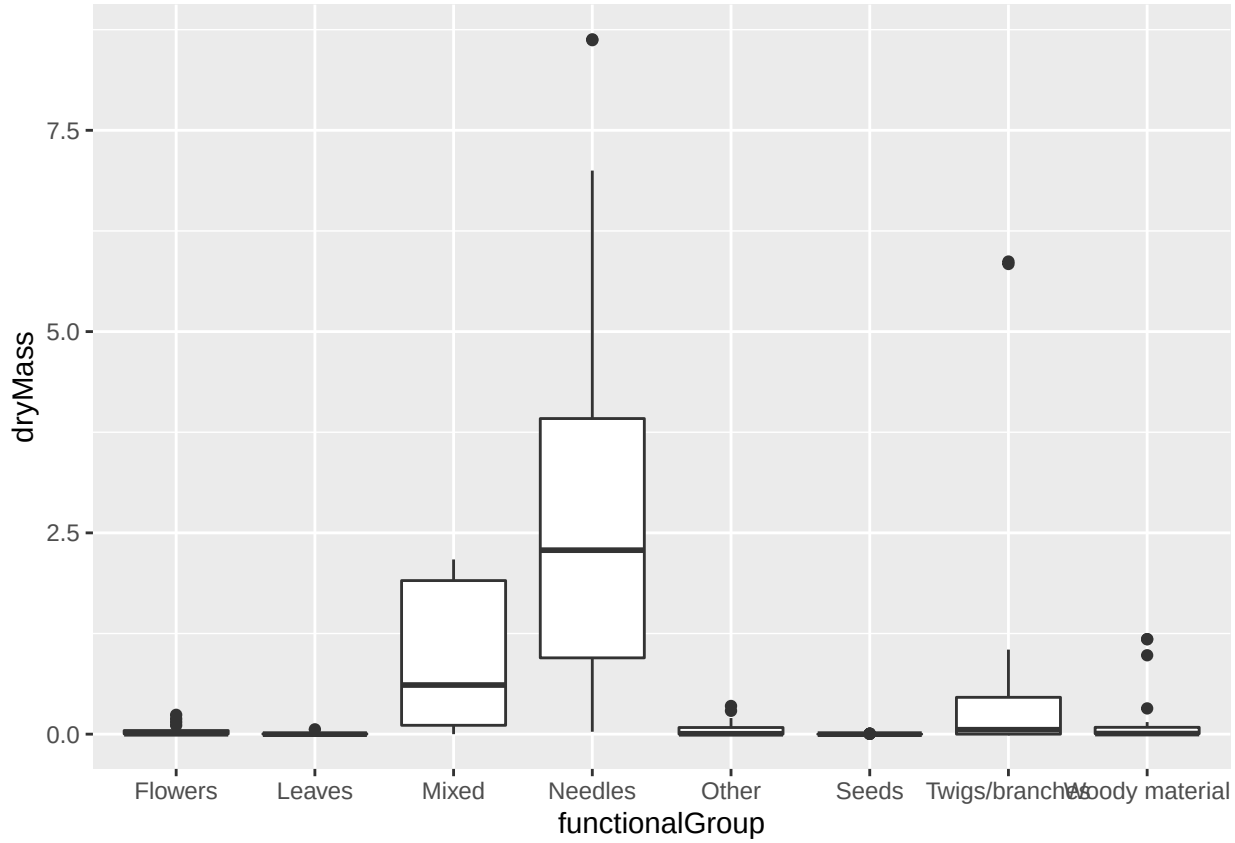
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(data = Litter, aes(x=Litter$functionalGroup)) +
  geom_bar()
```

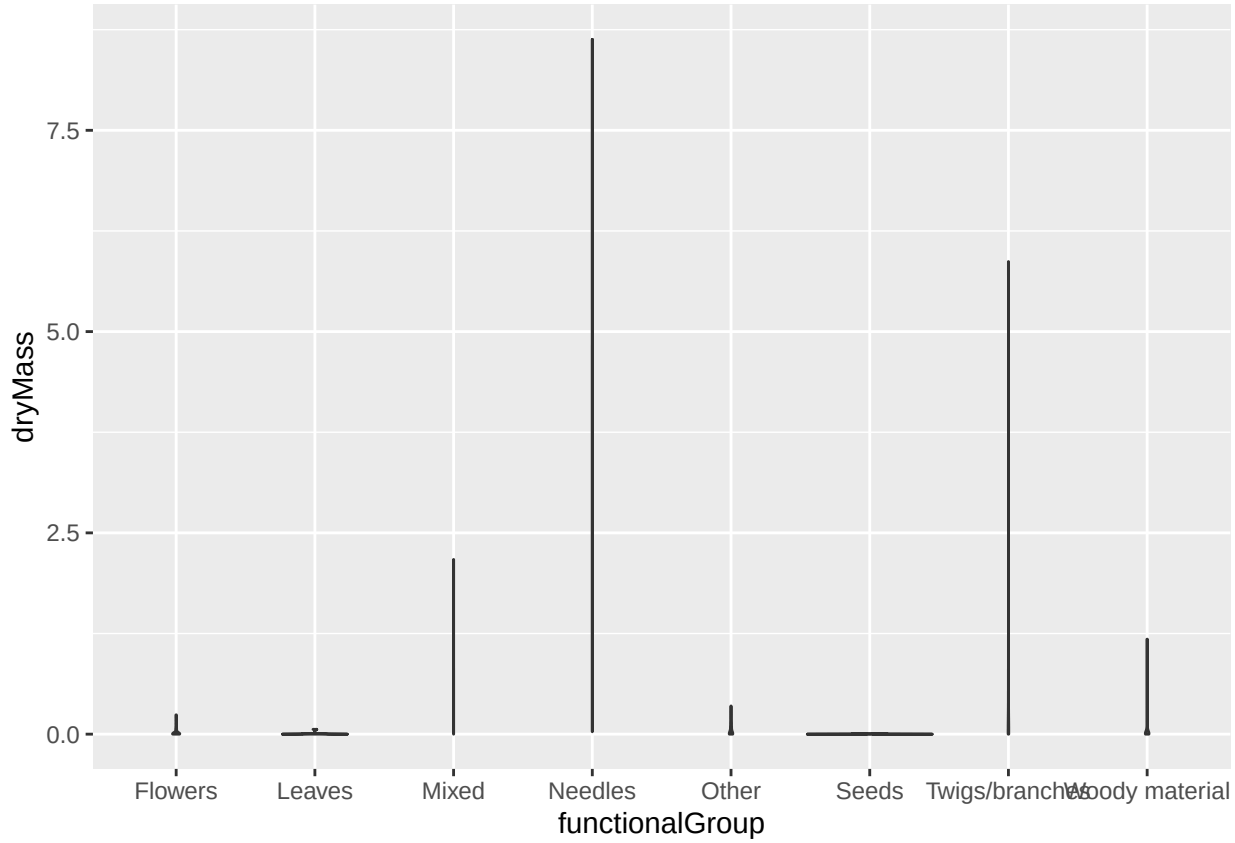


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




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



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

> Answer: The distributions of dry masses are centered close to zero, but there are some fairly high values so the plot is zoomed out. Essentially, the distributions and scale of the plot are such that we can't see much detail in the violin plots, so the boxplot works better overall for exploring the data.

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

> Answer: Based on the boxplot, needles and mixed litter appear to highest biomass at the sites.