Assignment 3: Data Exploration

Aye Nyein Thu

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.

```
#Loading necessary packages
library(tidyverse)
library(lubridate)
library(here)

#Checking the current working directory
getwd()
```

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

```
here()
```

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

```
#Uploading and naming datasets
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: Neocotinoids are one of the most widely used insectides in the United States, due to its effectiveness in treating insects and soil pects, and availabilty in the market at low cost (Nunez & Potter,2020). However, it poses a lot of harmful effects on ecosystems. First, it creates severe effects on pollinators, mainly to bees. Secondly, as it is in the systemic pesticides class, it reaches to all parts of plants and stay in the soil for a long time. When they are carried away by rain, it contaminates the waterbody and aquatic life. Finally, human are also victims of Neonicotinoids' side effects. It can cause tremors, low testosterone levels and even birth defects upon contact while pregnancy. Therefore, it is important to learn more about the ecotoxicology of neonicotinoids on insects.

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: With reference to the Forest Service, the U.S. Department of Agriculture, litter and woody debris are beneficial to both forest and aquatic ecosystems in many ways. They supports the natural nutrient cycling and plant growth, enhances soil quality, provides habitat for terrestrial and aquatic organisms, and reduces erosion and flooding. Therefore, it is worthwhile to study the litter and woody debris forms and the impacts they create for the environment.

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.Spatial Sampling Design: The sampling is executed in the tower plots. Trap placement within plots is either targeted or randomized, depending on the vegetation. 2.Temporal Sampling

Design: It collects samples on different time intervals. While group traps are sampled on a yearly basis, target sampling frequency for elevated traps may vary. 3. In 2018, there was sorting reductions in sampling and in 2020, the number of elevated traps were reduced without impacting estimates.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#Dimensions of 'Neonics' Dataset
dim(Neonics)
```

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

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...]

```
#Summary and Sorting the 'Effect'
summary_NeonicsEffect <- summary(Neonics$Effect)
sort(summary_NeonicsEffect, 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: As per the summary statistics, population and mortality are the most common effects that are studied while hormones and histology are the least common ones. Measuring population dynamics and mortality are important to analyzew if the impact of neonicotinoids on insects are huge and severe. Therefore, these are of particularlly interests to study compared to hormones.

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...]

```
#Finding the six mostly commonly studied species using 'maxsum' argument, 'summary' function
summary(Neonics$Species.Common.Name,maxsum=7)
```

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

Answer:The six most commonly studied species in the Neonics data set are Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee and Italian Honeybee. All of them are the pollinators which are essential for various plants and crops.

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

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

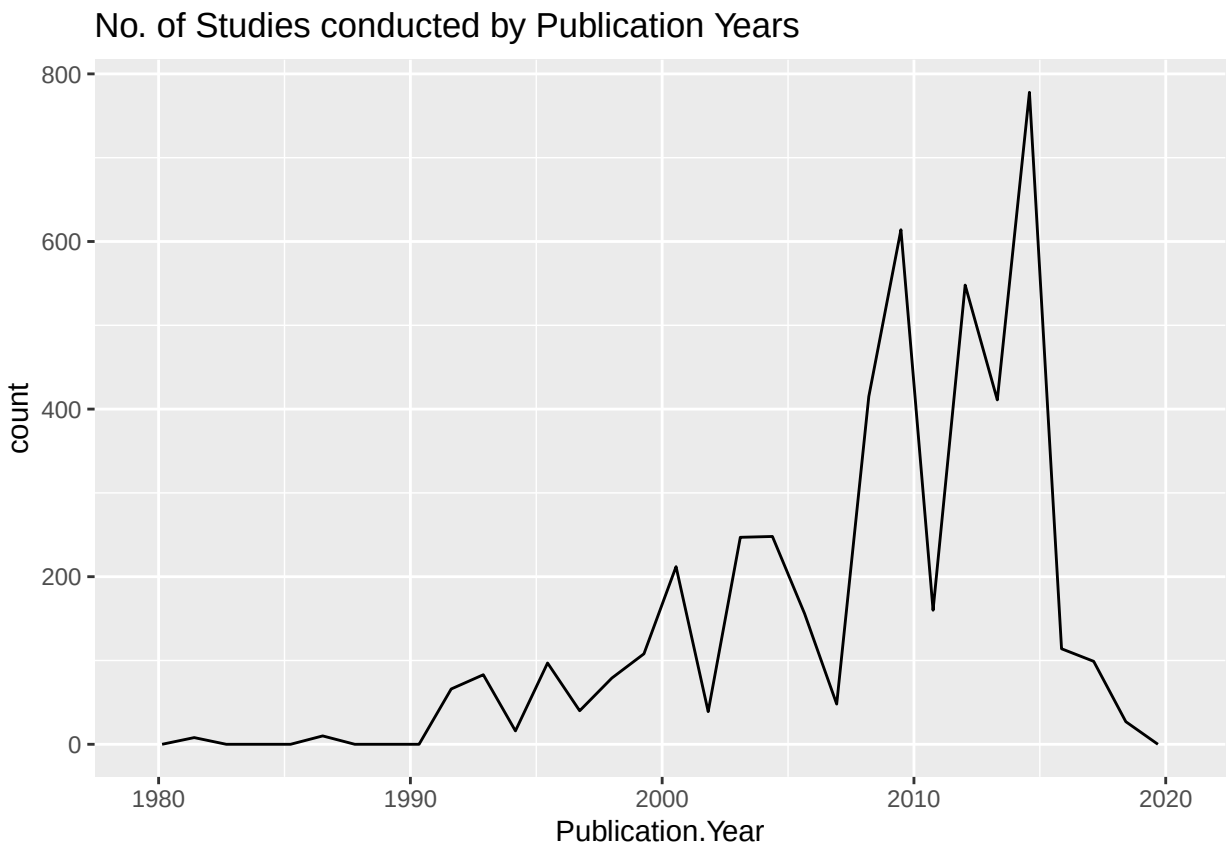
```
view(Neonics)
```

Answer: The class of "Conc.1..Author." is factor. It is not registed as numeric because it includes some mathematical symbols such as ~, /, >, < and NR.

Explore your data graphically (Neonics)

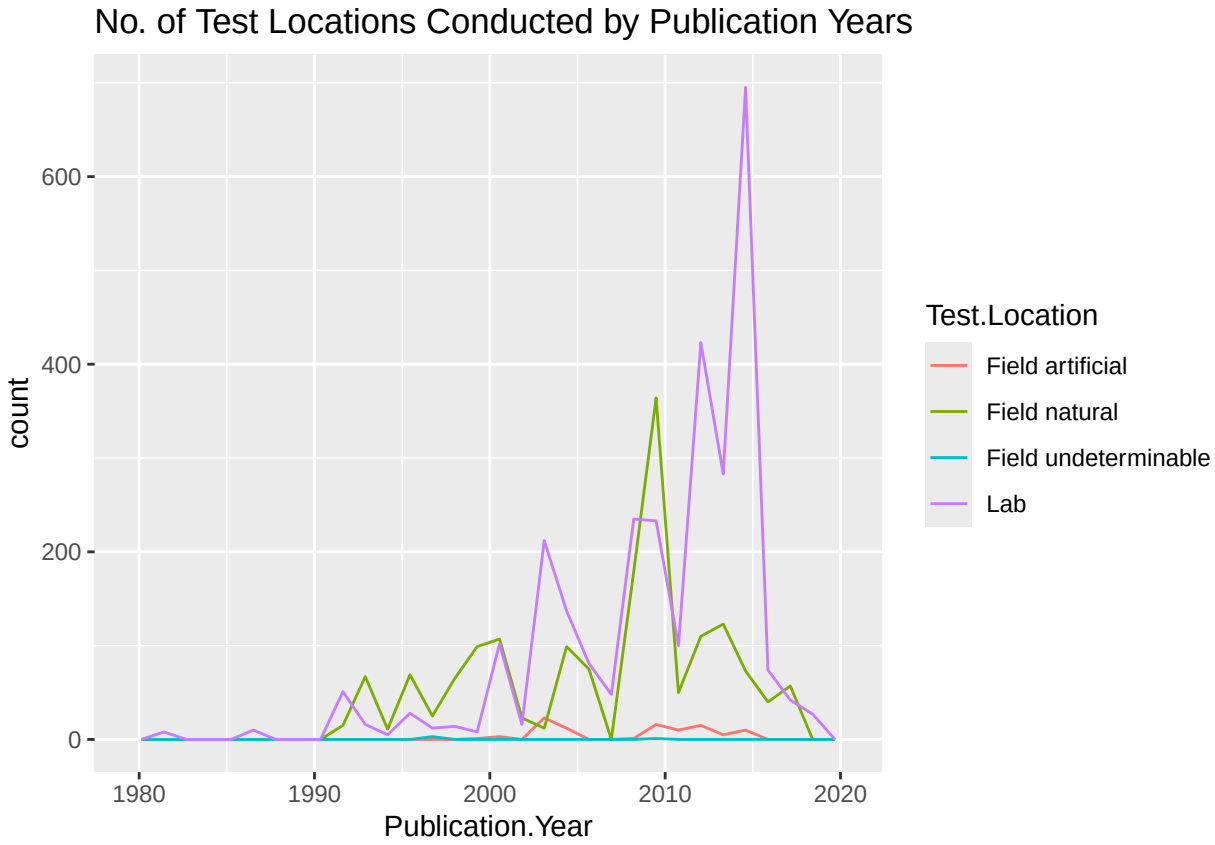
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
#Frequency Polygons: No. of Studies conducted by publication year
ggplot(Neonics) +
  geom_freqpoly(aes(x=Publication.Year)) +
  labs(title="No. of Studies conducted by Publication Years")
```



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

```
#Frequency Polygons: Test Locations conducted by publication year
ggplot(Neonics) +
  geom_freqpoly(aes(Publication.Year, color=Test.Location)) +
  labs(title="No. of Test Locations Conducted by Publication Years")
```



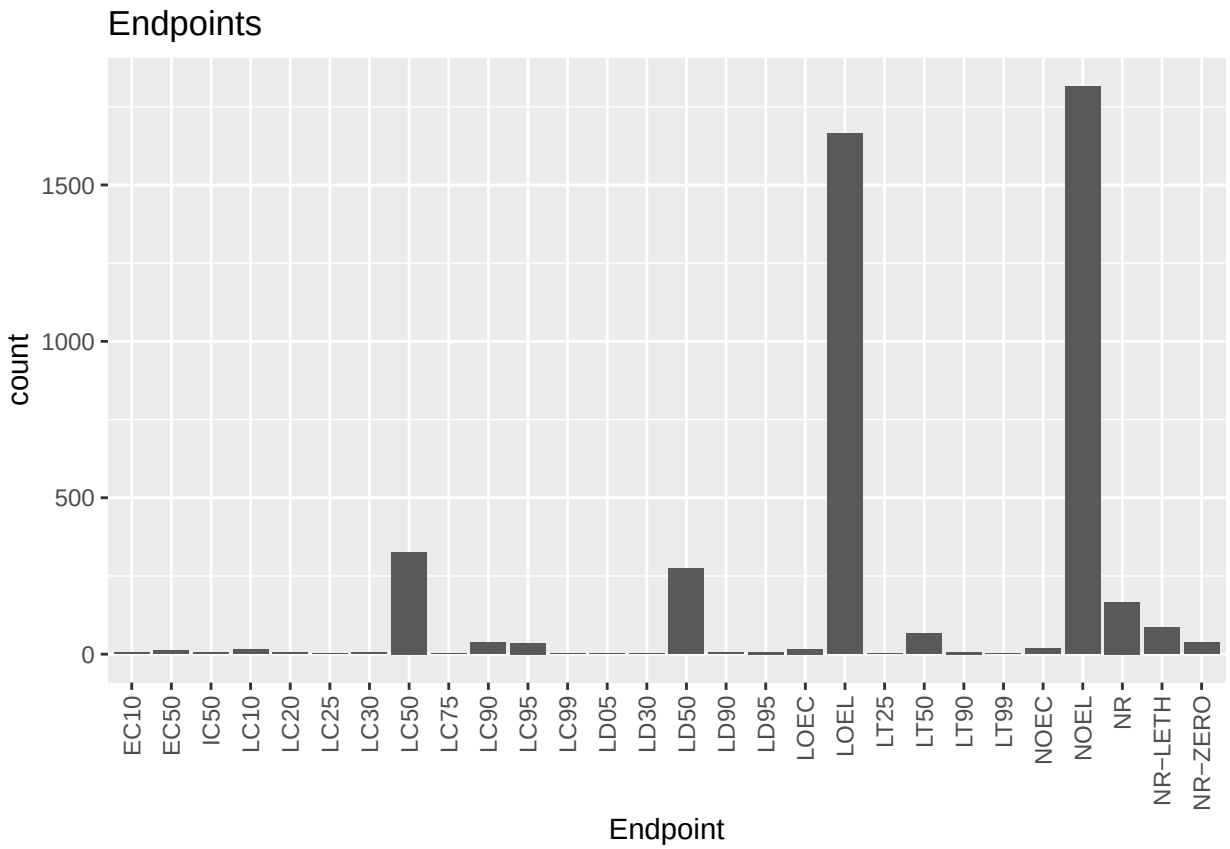
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The graph depicts that the most common test locations is lab, followed by field natural. The field artificial and the field undeterminable are low throughout the years. The test locations did not have much fluctuations from 1980 to 1990. Starting from 1990, the test natural and lab were showing the increasing trends but with high variations from years to 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...]

```
#Bar Graph: Endpoint
ggplot(Neonics) +
  geom_bar(aes(x=Endpoint)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  labs(title="Endpoints")
```



Answer:The two most common end points are LOEL and NOEL. LOEL is defined as Lowest-observable-effect-level, lowest dose (concentration) producing effects that were significantly different (as reportedby authors) from responses of controls (LOEAL/LOEC). NOEL 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.

```
#Determining class and changing from factor to date of 'collectDate'
class(Litter$collectDate)
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate, format="%y-%m-%d")
class(Litter$collectDate)
```

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

```r
unique(Litter$collectDate)
```

```
## [1] NA
```

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`?

```r
#Unique Vs. Summary of the Plot
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
```

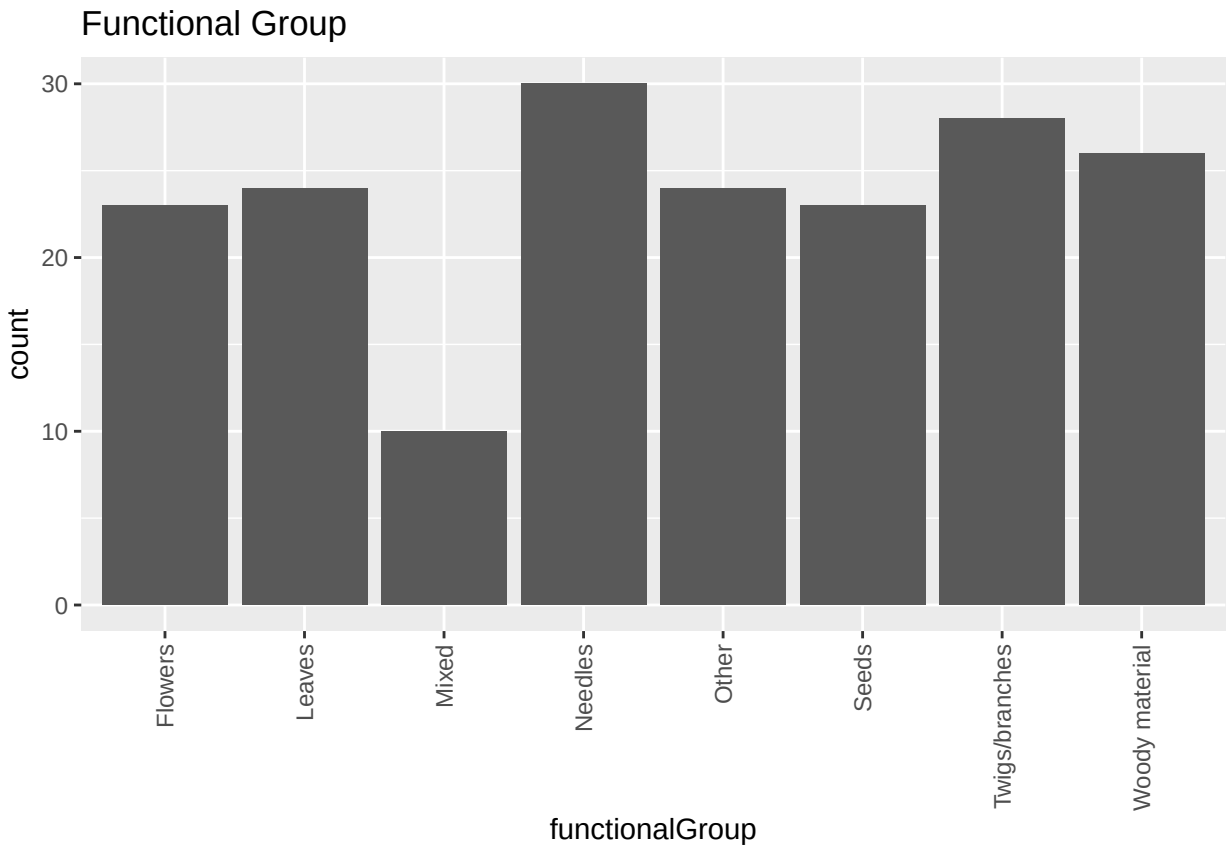
```r
summary(Litter$plotID)
```

```
## 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: A total of 12 different plots were sampled at Niwot Ridge. The information obtained from 'unique' function is the unique outputs/ types added to an object or variable. The 'summary' functions can tell both the types and counts of the object.

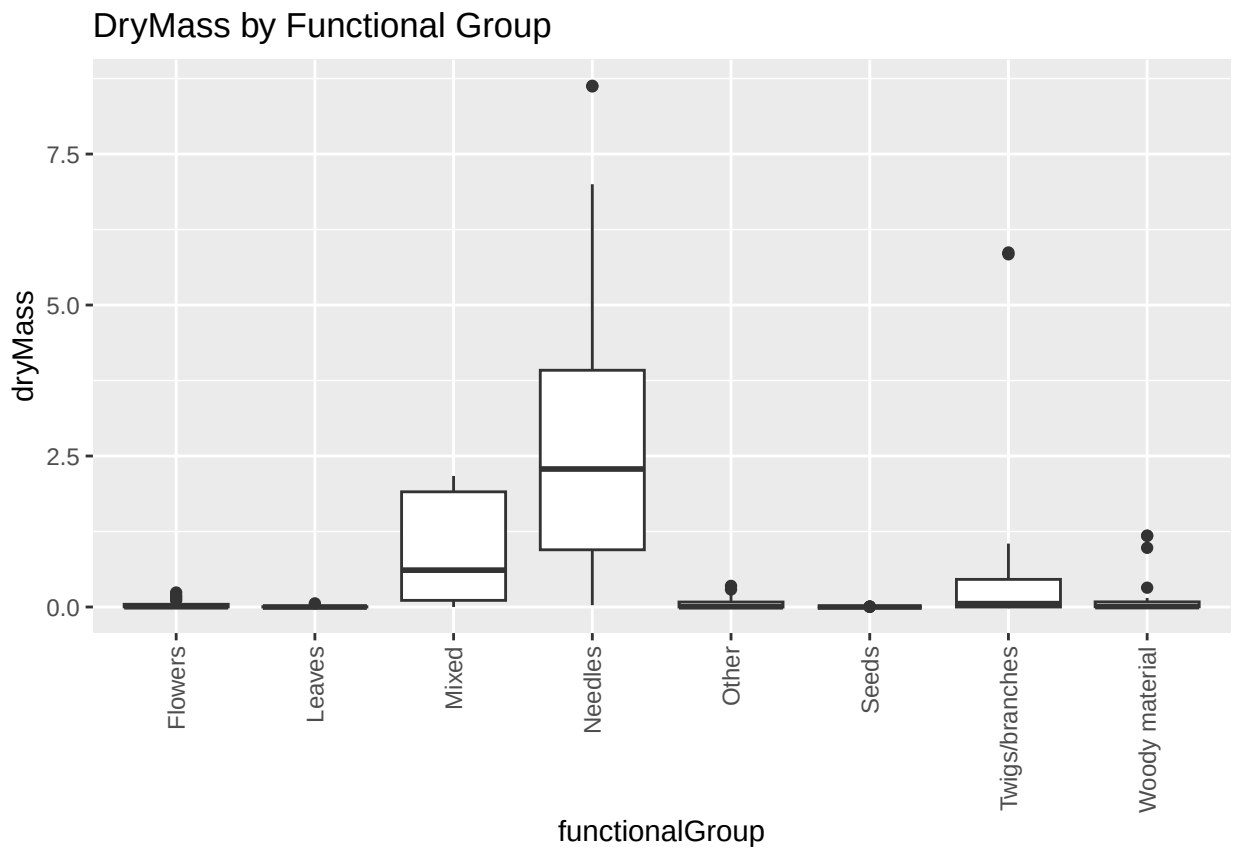
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
#Bar Graph: Functional Group
ggplot(Litter) +
  geom_bar(aes(x=functionalGroup))+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  labs(title="Functional Group")
```

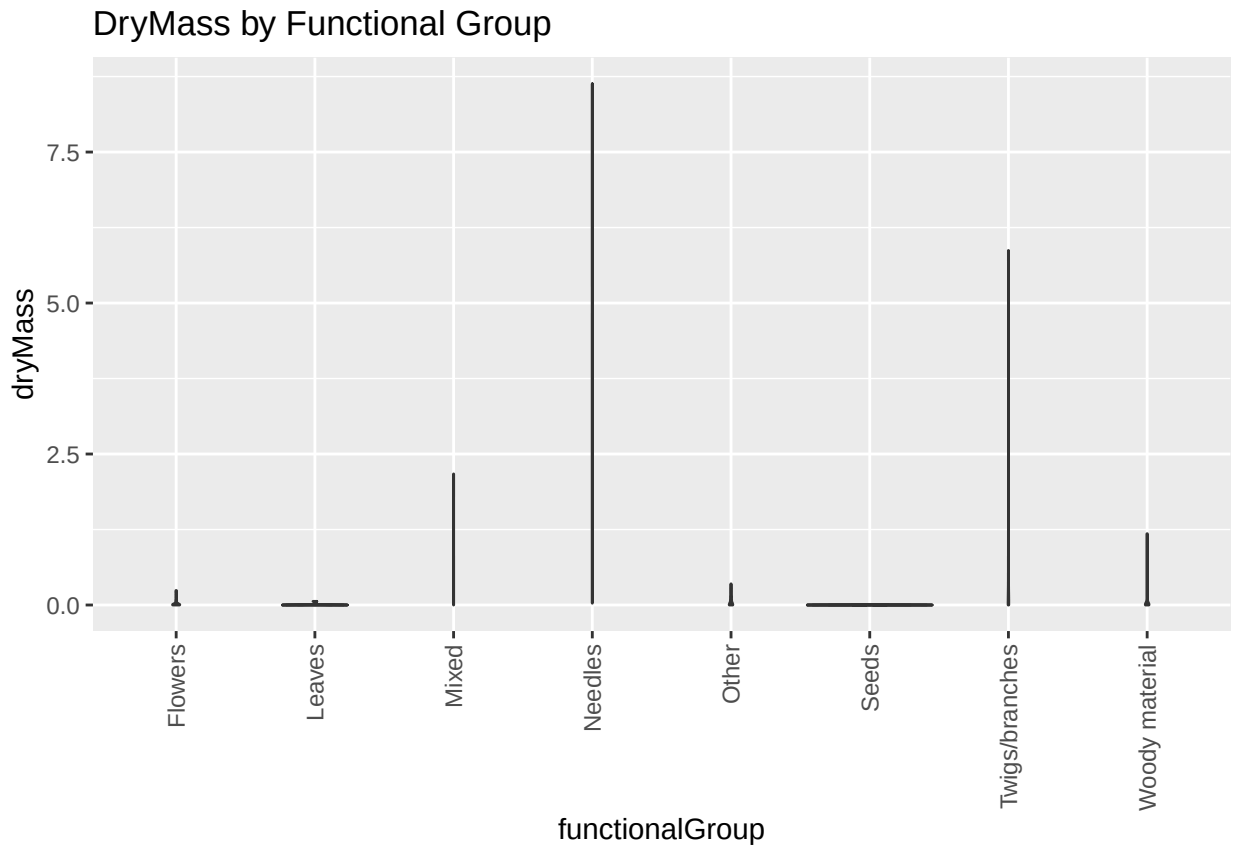


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#BoxPlot: DryMass by Functional Group
ggplot(Litter) +
  geom_boxplot(aes(x=functionalGroup, y=dryMass))+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  labs(title="DryMass by Functional Group")
```




```
#Violin Plot: DryMass by Functional Group
ggplot(Litter) +
  geom_violin(aes(x=functionalGroup, y=dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  labs(title="DryMass by Functional Group")
```



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

> Answer: In this case, the boxplot is more effective than the violin plot as it can be seen clearly that a majority of which functional group are equally distributed in terms of dryMass.

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

> Answer: Needles have the highest biomass at these sites.