Assignment 3: Data Exploration

Ying Liu

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. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.
6. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai.

The completed exercise is due on Sept 30th.

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. Be sure to include the subcommand to read strings in as factors.

```
getwd()
```

```
## [1] "C:/Users/Alina/Desktop/DUKE_22FALL/872/EDA-Fall2022/Assignments"
```

```
library(tidyverse)
library(lubridate)
Neonics <- read.csv("C:/Users/Alina/Desktop/DUKE_22FALL/872/EDA-Fall2022/Data/Raw/ECOTOX_Neonicotinoids_
Litter<-read.csv("C:/Users/Alina/Desktop/DUKE_22FALL/872/EDA-Fall2022/Data/Raw/NEON_NIWO_Litter_massdata
```

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: Though initially wide-praised for its low-toxicity to many beneficial insects and water water solubility, which allows them to be applied to soil and be taken up by plants. Now research found that neonicotinoid may be harmful for benificial insects like bees hrough low level contamination of nectar and pollen.It is important to understand the degree and mechanism of neonicotinoid on bees, and the subsequent influence in agriculture.

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:Litterfall and fine woody debris data may be used to es mate annual Aboveground Net Primary Produc vity (ANPP) and aboveground biomass at plot, site, and con nental scales. They also provide essential data for understanding vegetative carbon fluxes over time.

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: sampling occurs only in tower plots which are selected randomly within the 90% flux footprint of the primary and secondary airsheds. 1. plot edges must be separated by a distance 150% of one edge of the plot 2. plot centers must be greater than 50m from large paved roads and plot edges must be 10m from twotrack dirt roads 3. plot centers must be 50m from buildings and other non-NEON infrastructure

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 on the "Effect" column, 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 are population, mortality, behavior,feeding behavior.because such indexes directly reflect the influence of neonicotinoid on insects lifespan, and these features can directly lead to problems in agriculture

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.

```
#summarize the column, and then rank the top 6
sort(summary(Neonics$Species.Common.Name), decreasing=T)[1:6]
```

```
##                (Other)             Honey Bee        Parasitic Wasp
##                    670                   667                   285
## Buff Tailed Bumblebee   Carniolan Honey Bee            Bumble Bee
##                    183                   152                   140
```

Answer:Honey Bee,Parasitic Wasp, Buff Tailed Bumblebee,Carniolan Honey Bee,Bumble Bee Most are bees and all are benificial insects that are viral to the sustainable development of environment

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

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

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

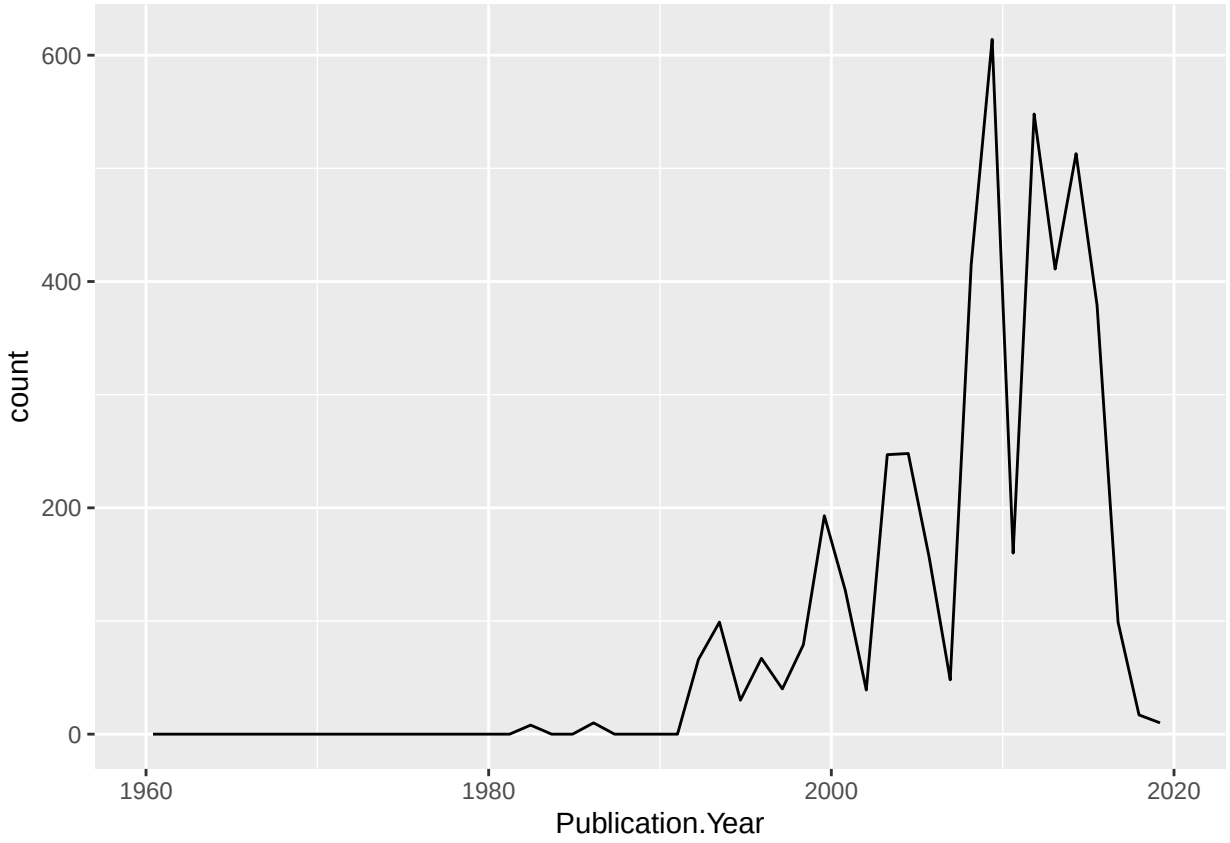
Answer:It's factor.Because some of they are range, not specific number

Explore your data graphically (Neonics)

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

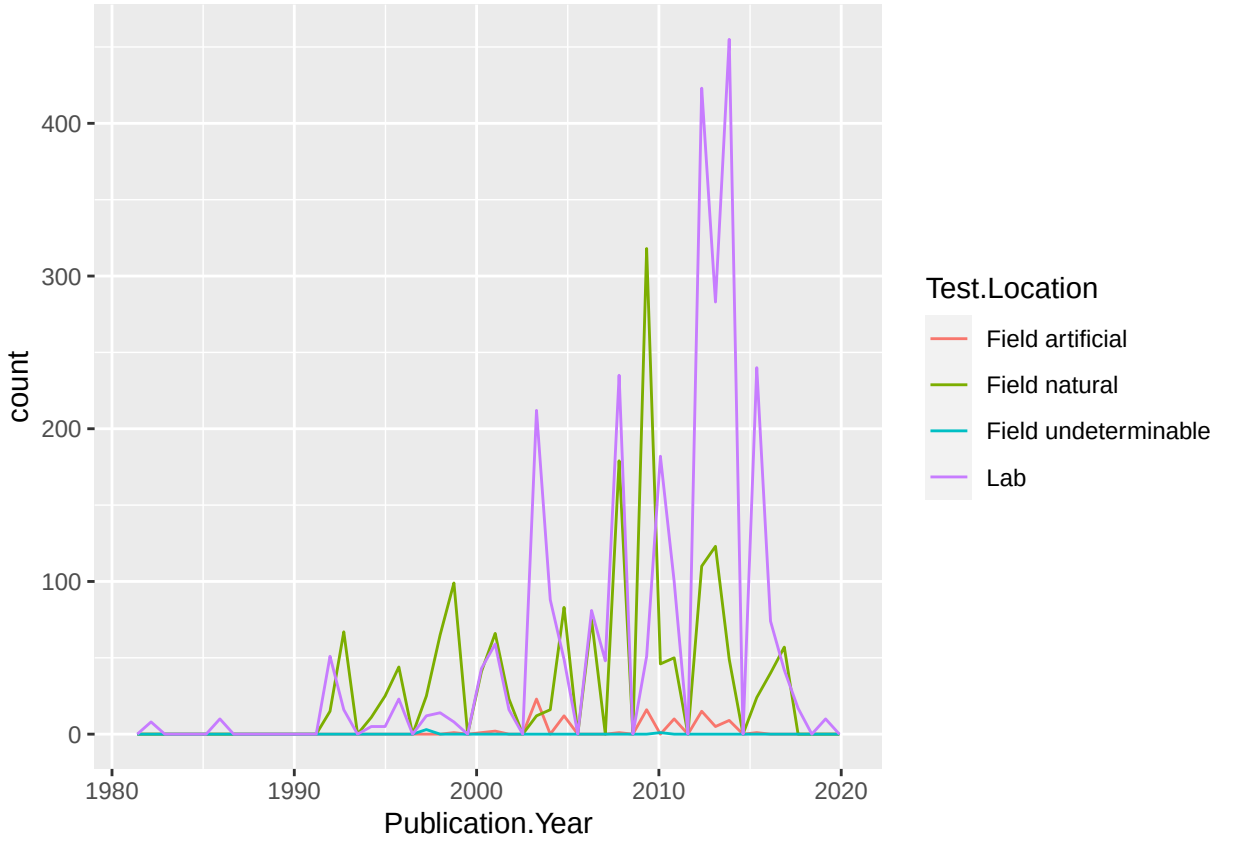
```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50)+
  scale_x_continuous(limits = c(1960, 2020))
```

```
## Warning: Removed 3 row(s) containing missing values (geom_path).
```



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

```
ggplot(Neonics, aes(x = Publication.Year, colour = Test.Location)) +
  geom_freqpoly(bins = 50)
```




```
  scale_x_continuous(limits = c(1960, 2020))
```

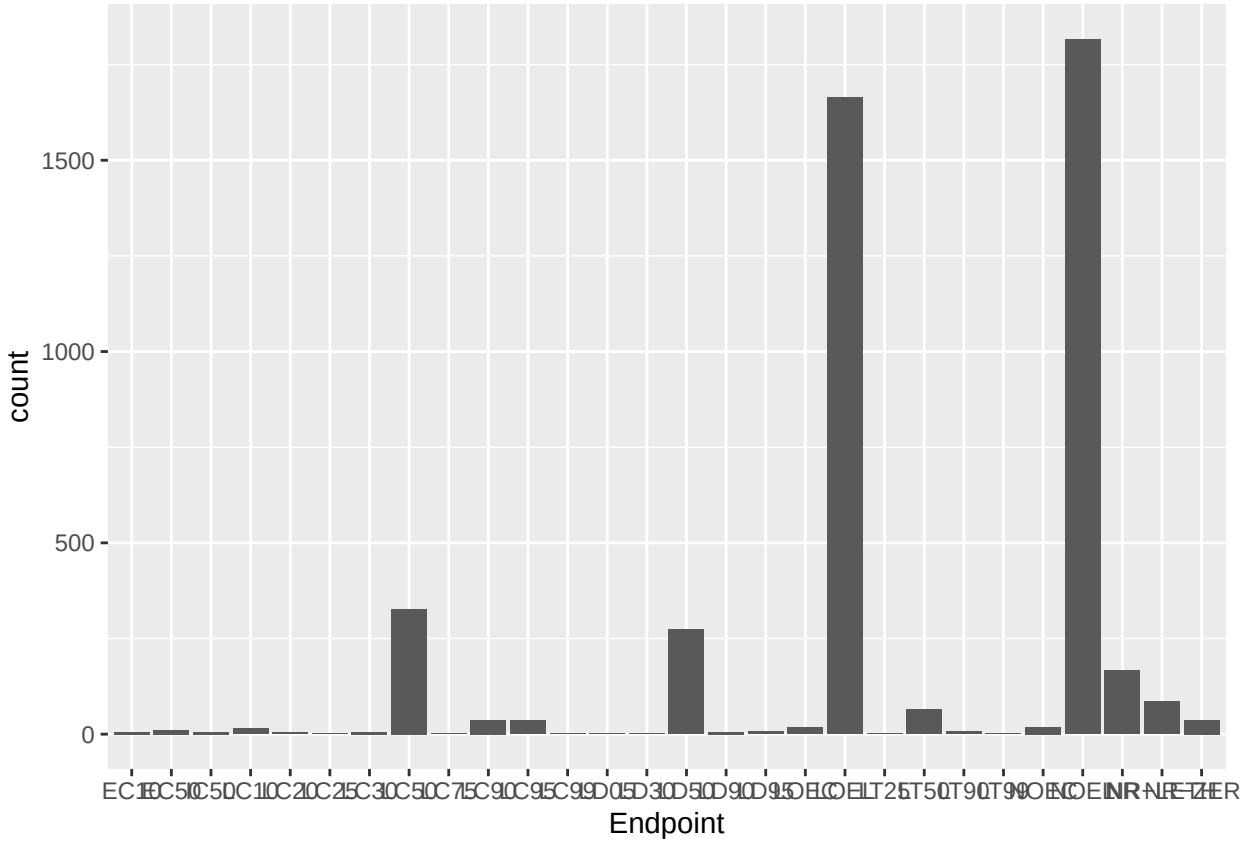
```
## <ScaleContinuousPosition>
##  Range:
##  Limits: 1.96e+03 -- 2.02e+03
```

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

> Answer:the most common is lab. as time goes by, more tests are conducted under lab and field natural

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(Neonics, aes(x = Endpoint)) +
  geom_bar()
```



Answer:LOEL,NOEL are the most common endpoints. LOEL(Lowest-observable-effect-level): lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC) NOEL(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.

```
class(Litter$collectDate)
```

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

```
#it is factor
Litter$collectDate<-ymd(Litter$collectDate)
#confirm data type has changed to date
class(Litter$collectDate)
```

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

```
#class(Litter$collectDate) = "date"
unique(Litter$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`?

```
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
```

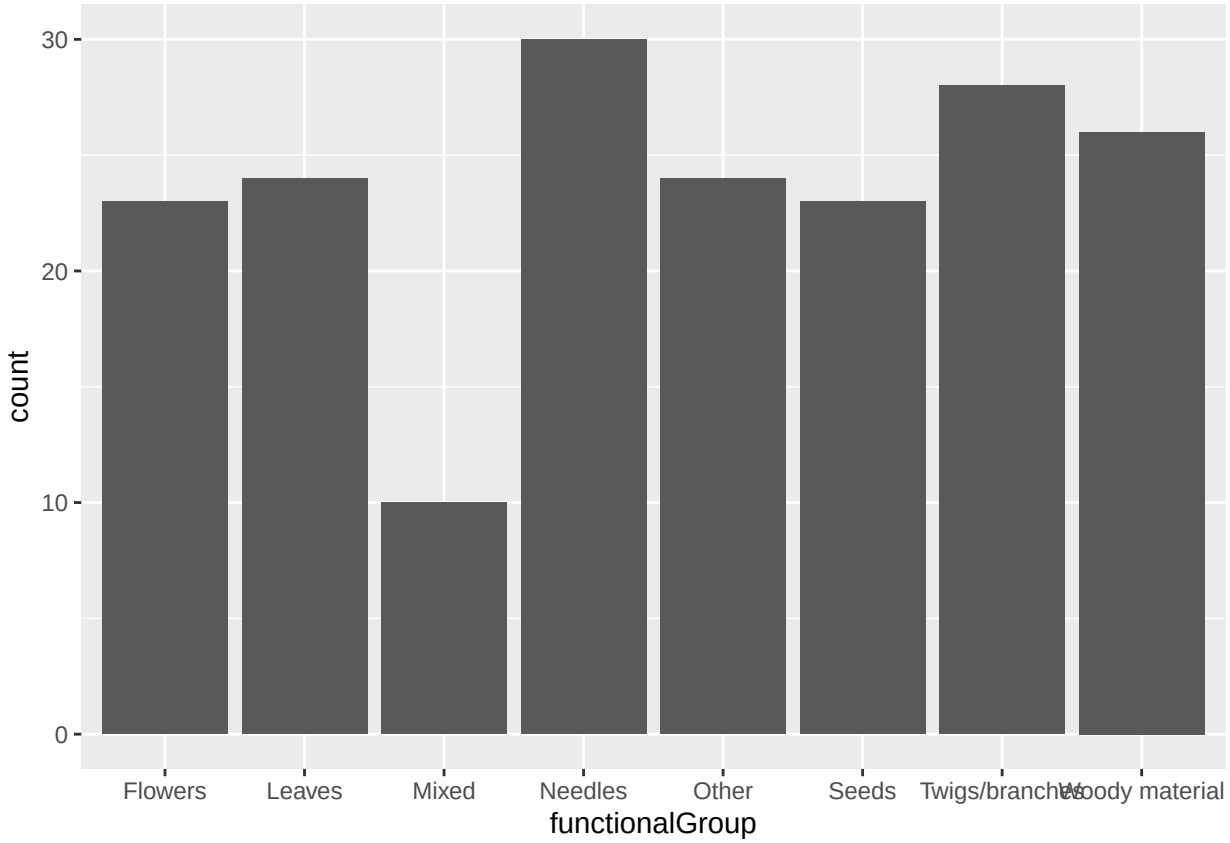
```
summary(Litter$plotID)
```

```
## 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 unique shows the different values summary can not only show different values, but can count the numbers of each different values

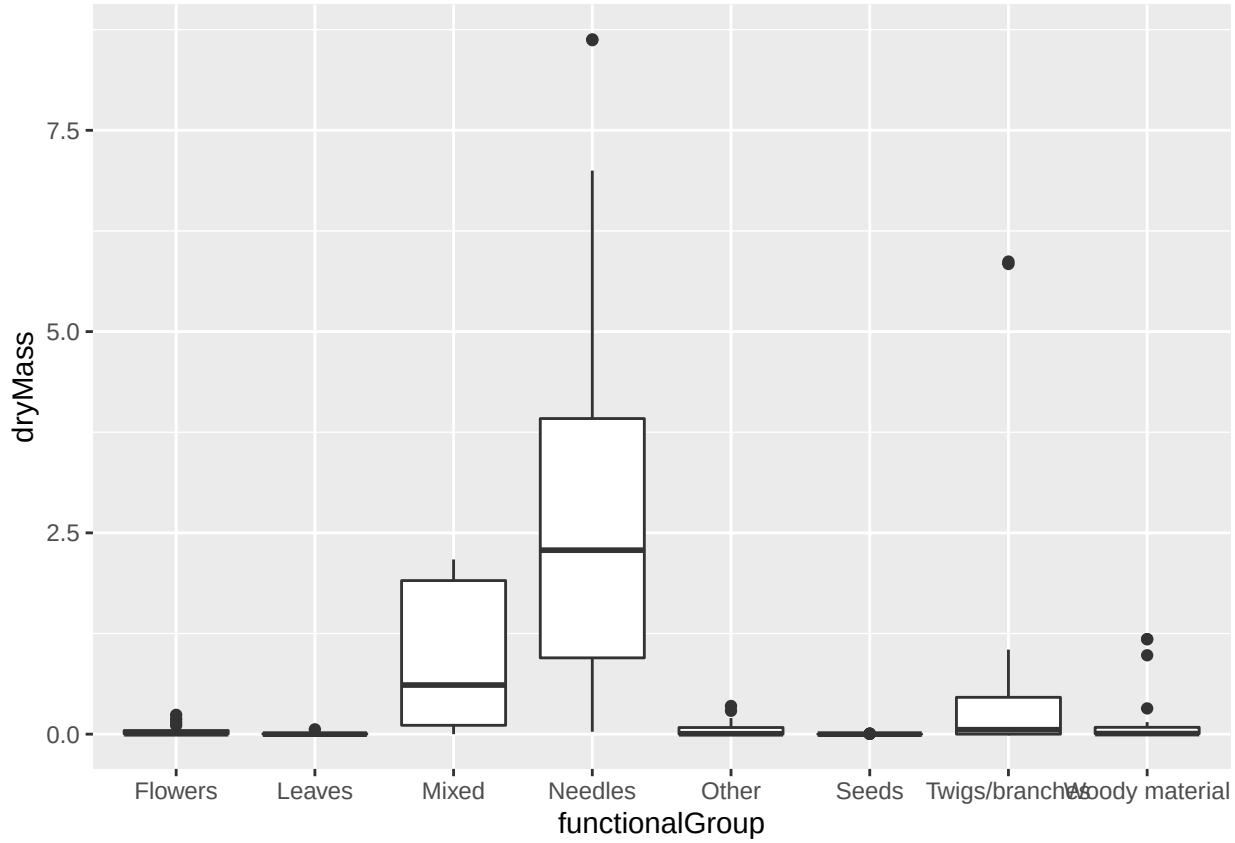
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
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.

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

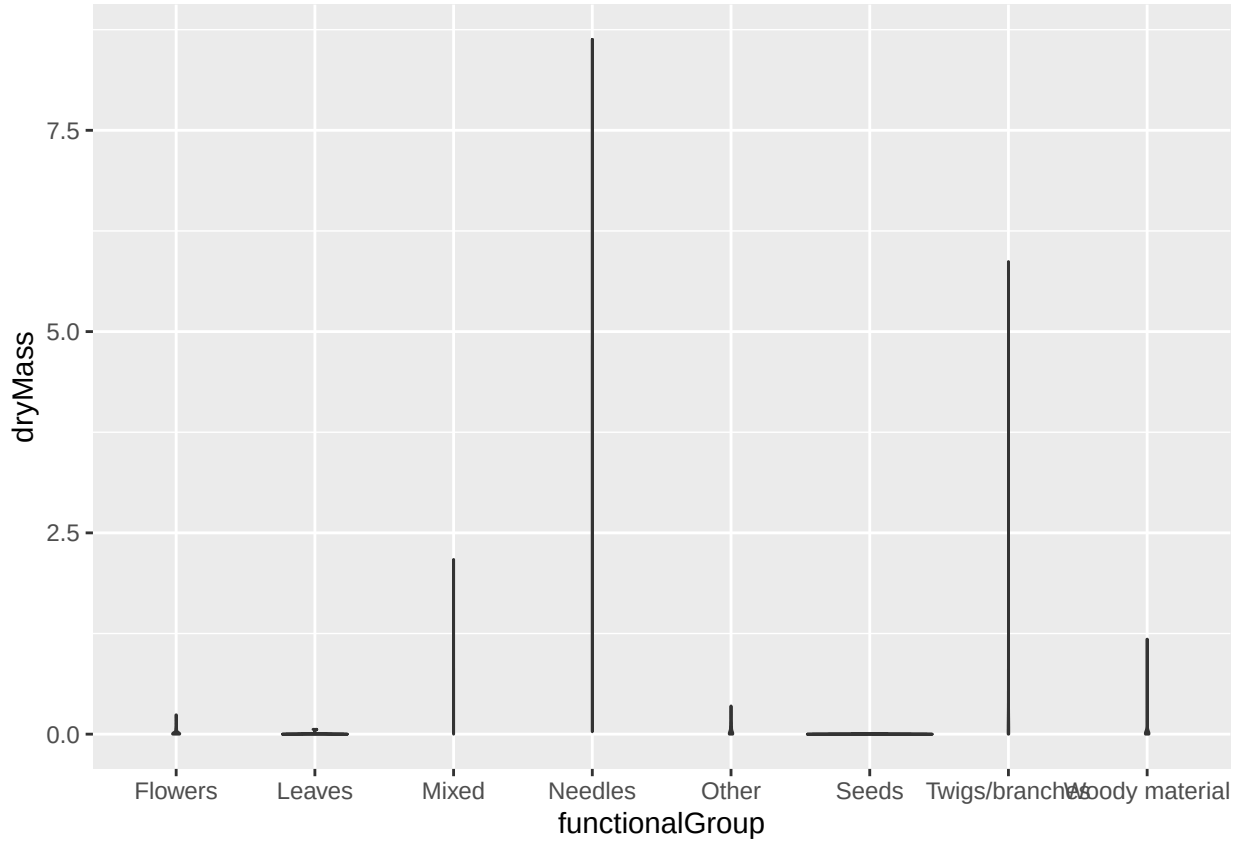



```
#violin plot
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup , y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75))
```

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```



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

> Answer:Both plots can show the descriptive stastics of the numeric data. Violin plot combines a box plot and data density.But in this case box plot is better because it's more clear and evenly distributed on plotting area. Since most drymass data are small, tehy tend to culster in the bottom of violin, making the plot difficult to read.

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

> Answer:needles