Assignment 3: Data Exploration

Soumya Mathew

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.

```
library(formatR) #loading package for formating

knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE) #trying to wrap codes in the pd

#Question 1
getwd() #checking working directory
```

```
## [1] "C:/Users/user/Desktop/Soumya/Year 2/EDA/EDA-Fall2022/Assignments"
```

```
#install.packages("tidyverse") #installing package

#library(tidyverse) #loading package

#uploading dataset1
Neonics <- read.csv("C:/Users/user/Desktop/Soumya/Year 2/EDA/EDA-Fall2022/Data/Raw/ECOTOX_Neonicotinoid

#uploading dataset2
Litter <- read.csv("C:/Users/user/Desktop/Soumya/Year 2/EDA/EDA-Fall2022/Data/Raw/NEON_NIWO_Litter_mass
```

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: Neonicotinoids are used insecticides to get rid of pests that attacked specific kinds of agricultural crops. However, plants absorb the chemical substances, which then remains to their leaves and pollen. Neonicotinoids also have an impact on non-target species and crucial pollinators for agriculture. The ecotoxicology of neonicotinoids on all insect groups must therefore be examined.

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: As a source of energy for aquatic ecosystems, a habitat for terrestrial and aquatic organisms, and a contributor to structure and roughness, which influences water flows and sediment transport, woody debris and litter play important roles in carbon budgets and nutrient cycling in forest and stream ecosystems.

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.To gather samples of woody debris, ground traps were deployed in accordance with the appropriate procedure. 2.By calculating their elemental analysis, the carbon, nitrogen, and lignin content of litterfall and debris is taken into account. 3. The height of the woody vegetation and plot sizes are used to determine the best sites for collecting litterfall samples.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)  #checking dimensions
```

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

```
dim(Litter)  #checking dimensions
```

```
## [1] 188  19
```

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: Population, Mortality, Behaviour are the top 3 effects of insectisides on insects. Since it reveals how the chemical affects the insect group overall, it is crucial to research the ecotoxicology of the chemicals on these insects.It shows the consequences of over using insecticides. Additionally, it aids in the conservation of non-target species by identifying the effects of the active substances on those that should be prevented.

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.

```
summary(Neonics$Species.Common.Name, 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 results show that Bees are the most commonly studied species in the dataset. We depend on bees and other pollinators every day; bees are among the most significant pollinators of food crops in the world. In fact, pollination is necessary for one out of every three mouthful we take. However, because of the widespread use of pesticides and other environmental variables, such as climate change, bee populations continue to drop.

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.)  #checking class
```

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

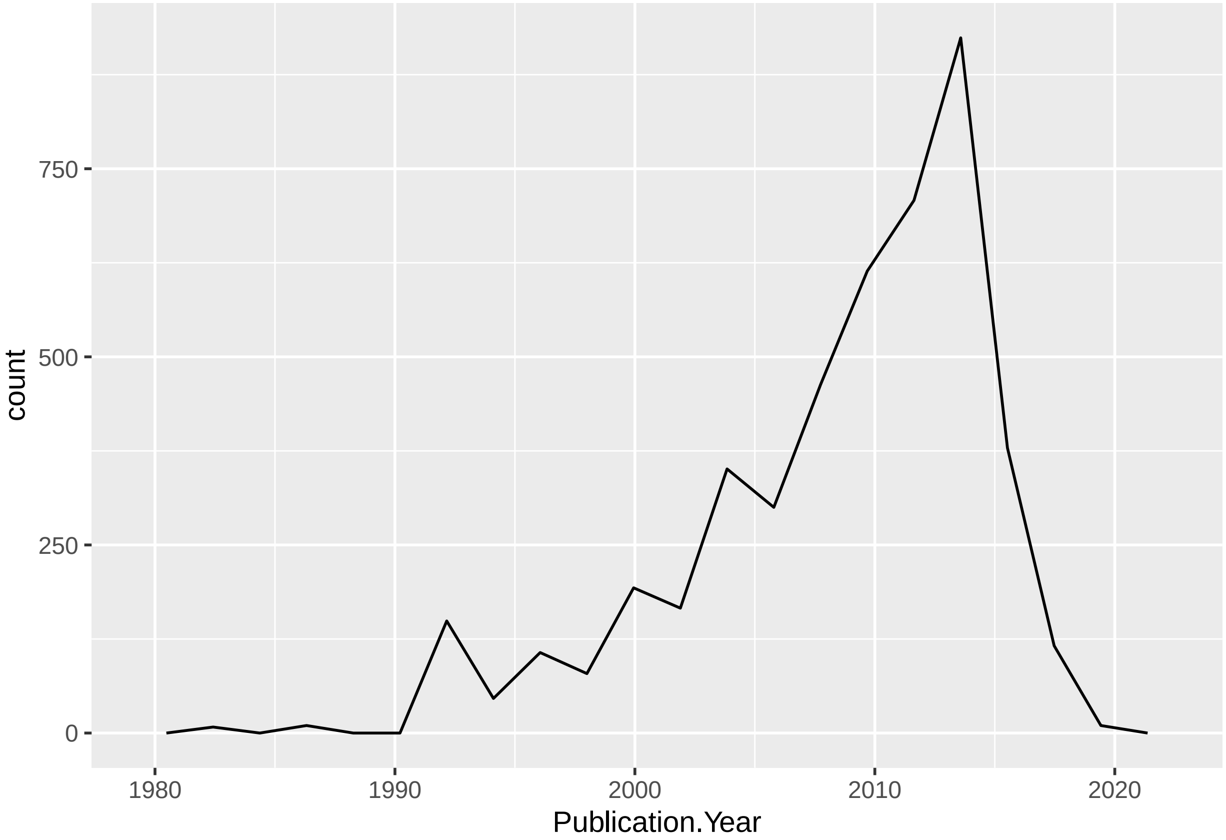
Answer: Since we imported the dataset using the subcommand "stringsAsFactors = TRUE," As a result, the information in the Conc.1..Author column was transformed into factors.

Explore your data graphically (Neonics)

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

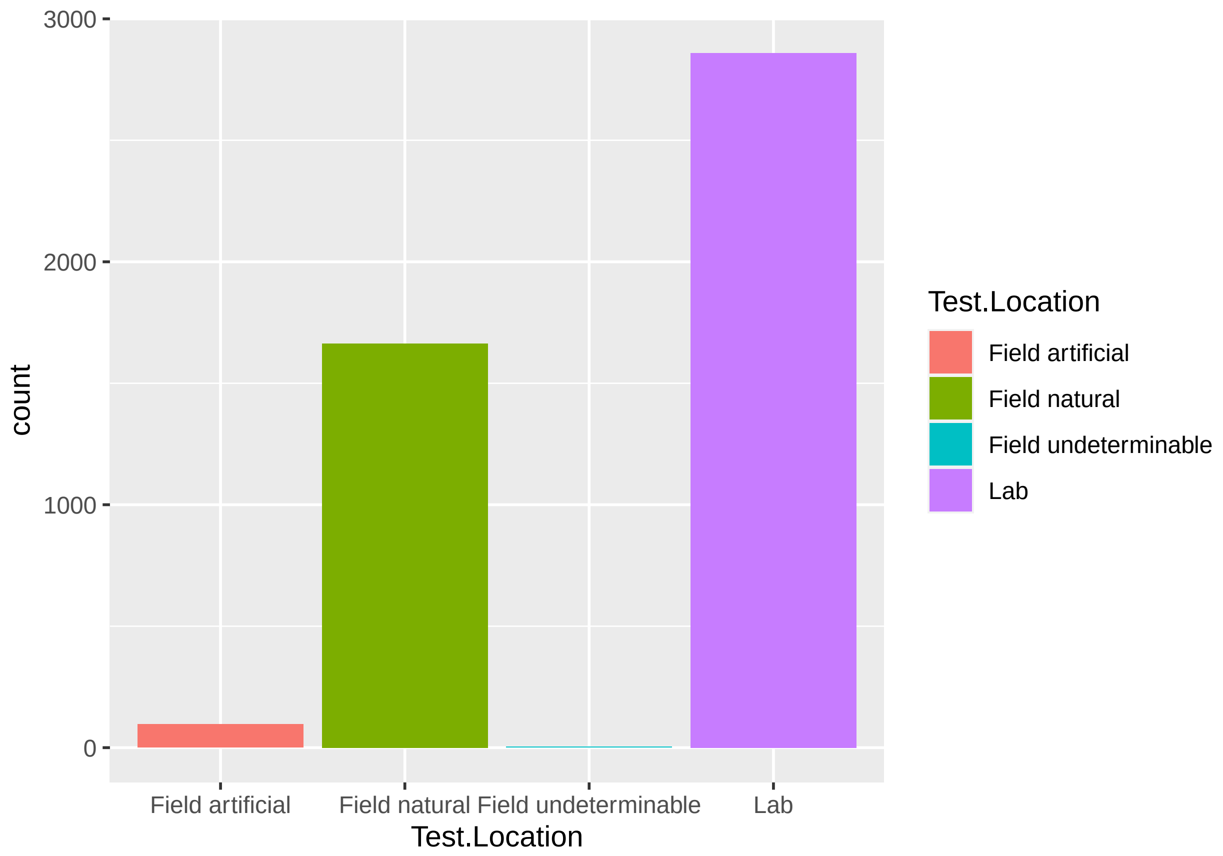
```
library(ggplot2)  #loading packages

ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year), bins = 20)
```



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 = Test.Location, fill = Test.Location)) +
    geom_bar()
```

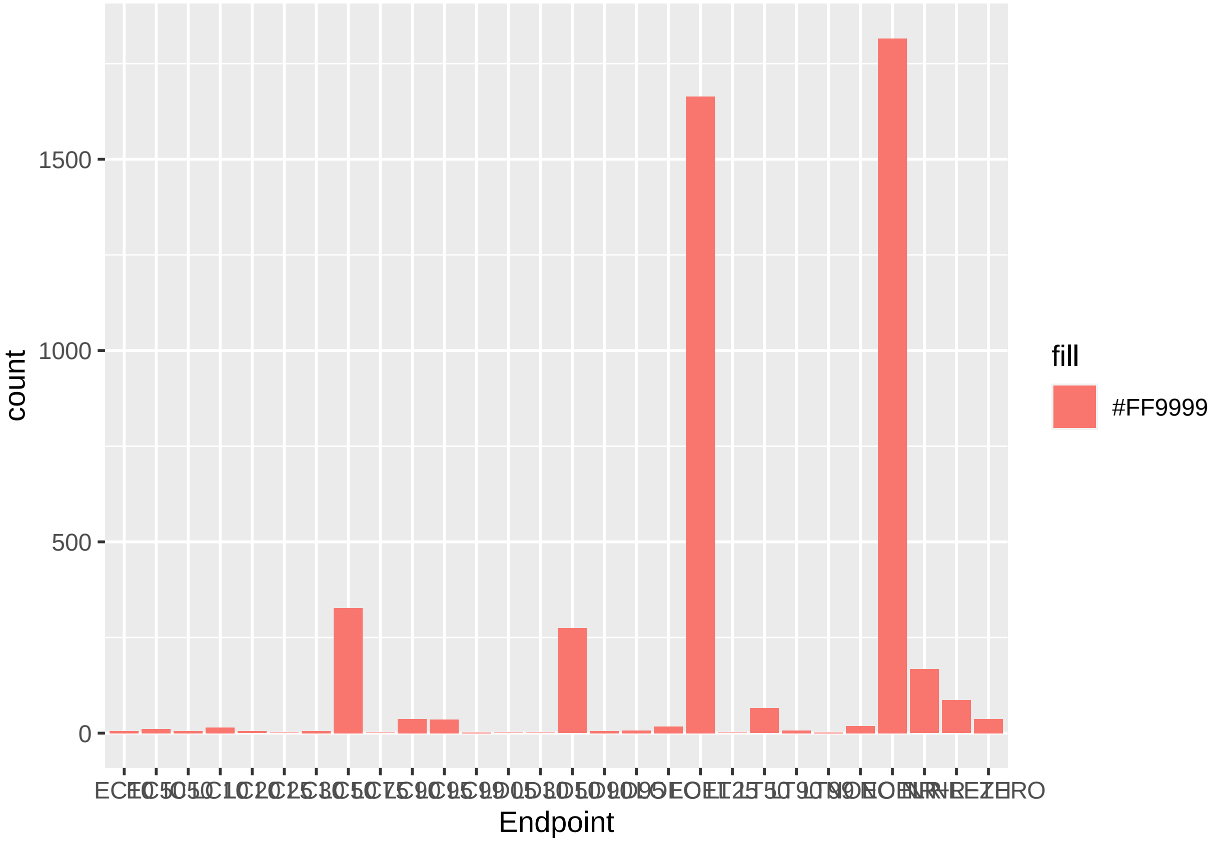


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: We infer from the graph that the most typical test locations are the lab and the field. Unlike a lab, a field organically changes over time.

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) + geom_bar(aes(x = Endpoint, fill = "#FF9999"))
```



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

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)  #determining class -> not a date, its factor
```

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

```
head(Litter$collectDate, 5)  #checking the current format of top 5 rows
```

```
## [1] 2018-08-02 2018-08-02 2018-08-02 2018-08-02 2018-08-02
## Levels: 2018-08-02 2018-08-30
```

```
library(lubridate)  # loading package
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
datesasfactor <- (Litter$collectDate)

date_format <- c(ymd(datesasfactor))  #changing format to Date

class(date_format)  #checking class of data column
```

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

```
unique(date_format)  # unique dates used in Litter are 2nd August 2018 and 30th August 2018
```

```
## [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
```

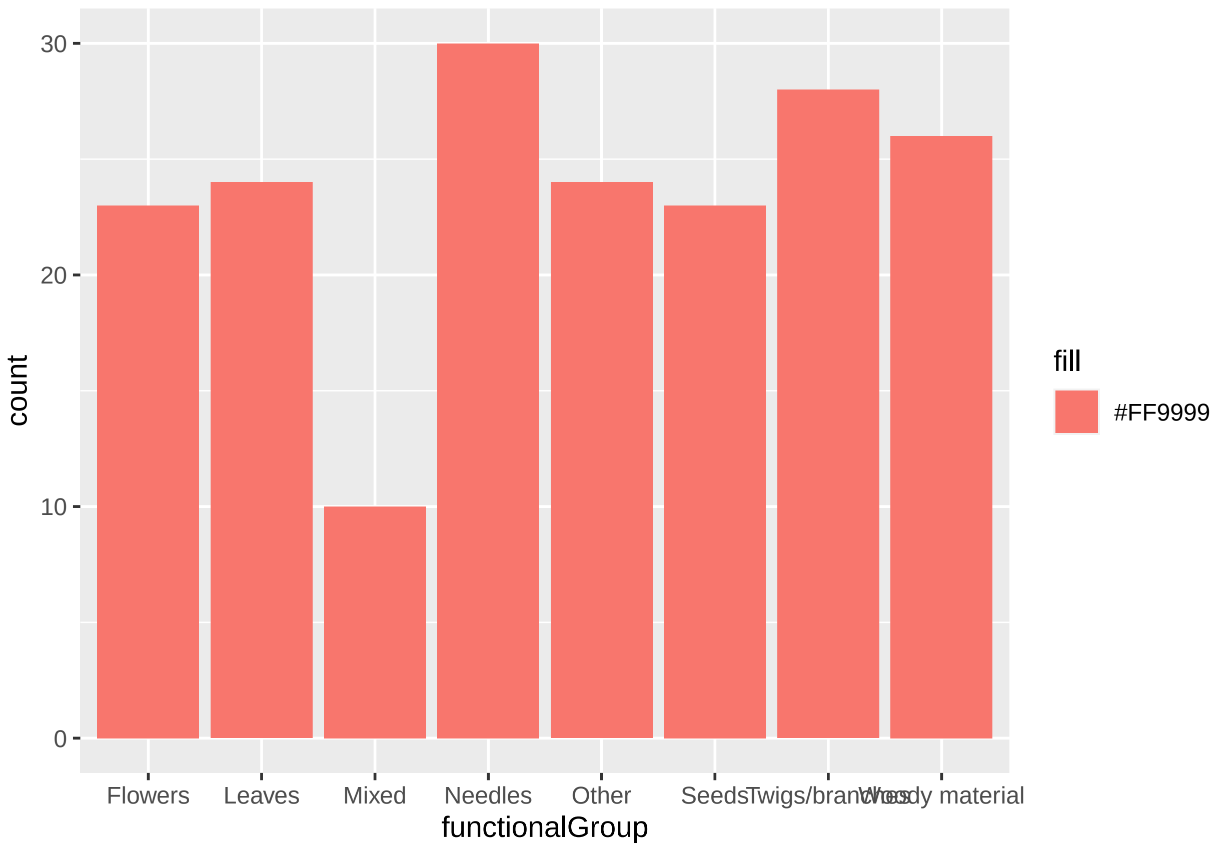
```
length(unique(Litter$plotID))  # out of 188 total plots 12 were unique
```

```
## [1] 12
```

Answer: The unique()function displays the dataset's unique values after removing duplicate entries. While summary() provides an in-depth overview of the dataset's findings (ket statistics by column).

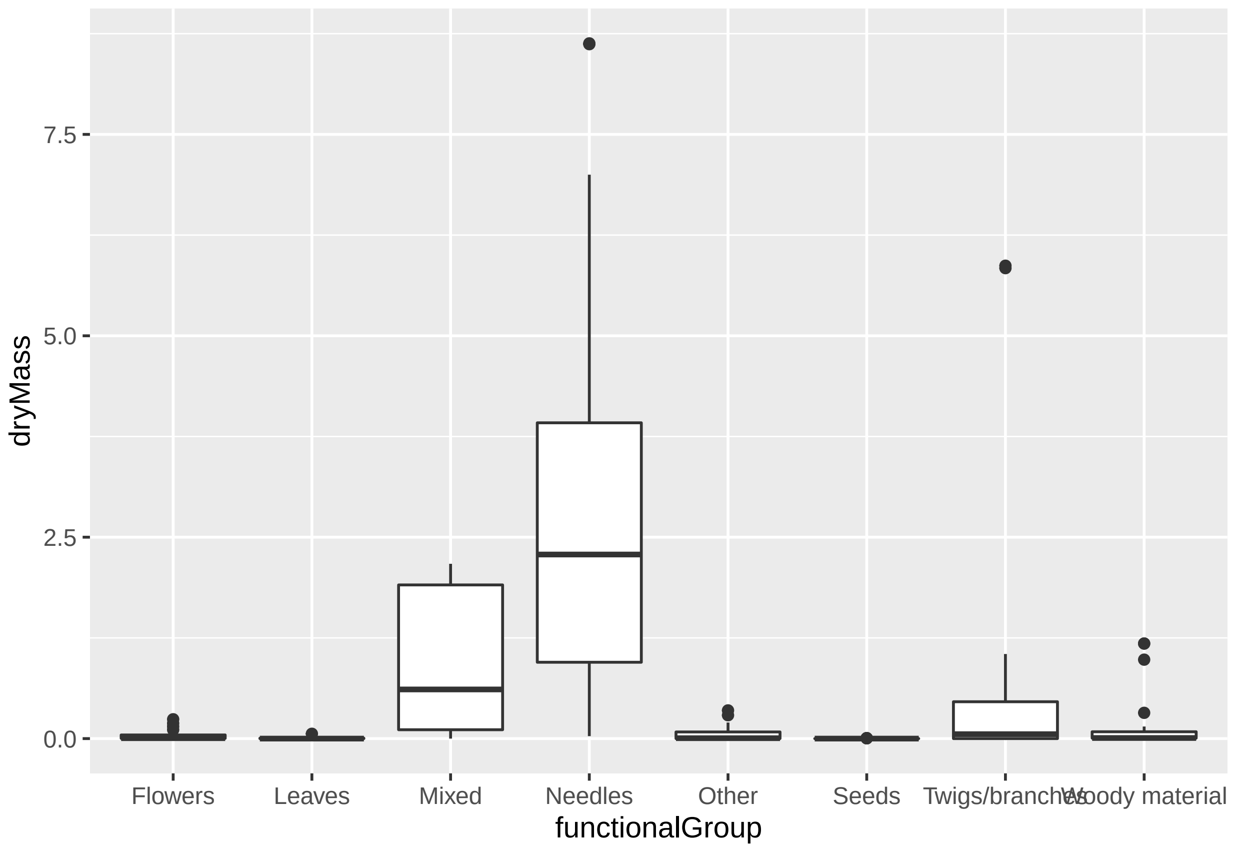
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(x = functionalGroup, fill = "#FF9999")) +
    geom_bar()
```



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

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

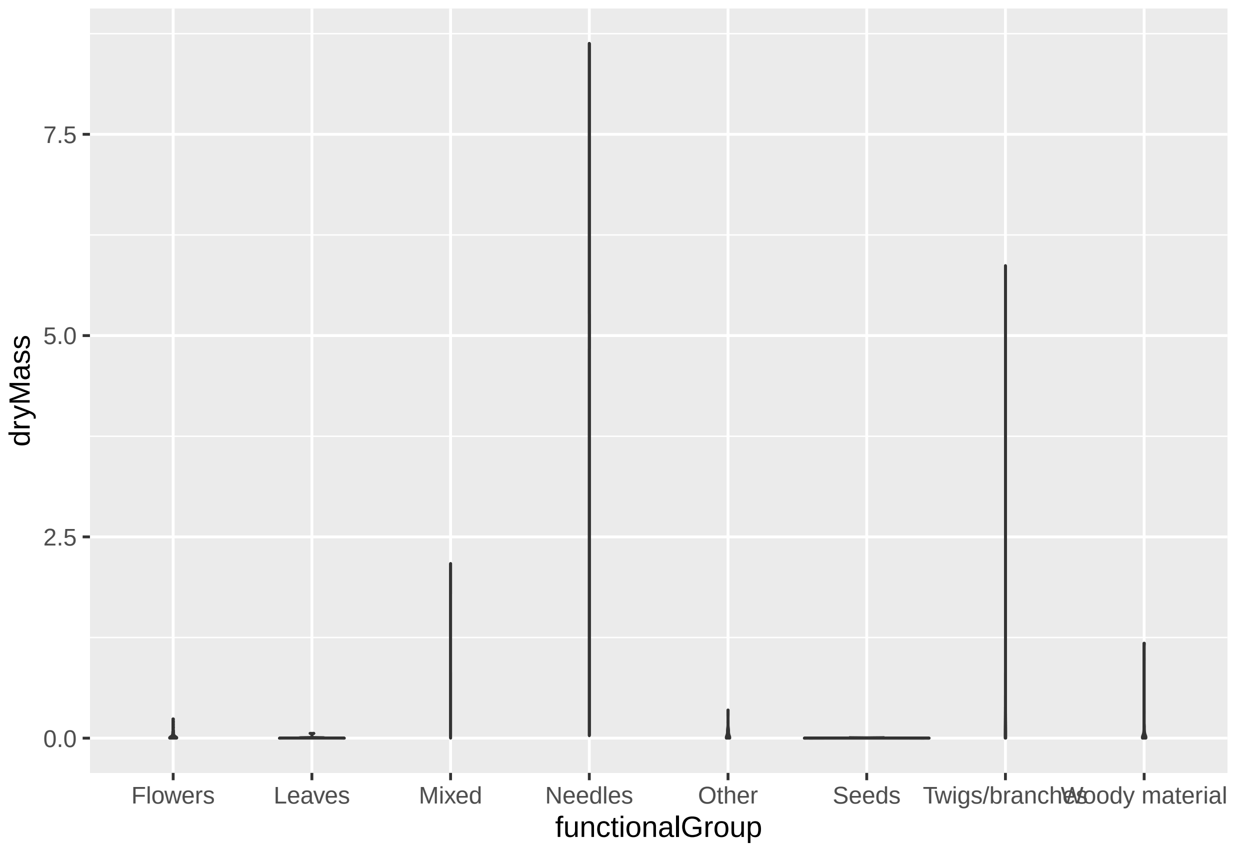


```
ggplot(Litter) + geom_violin(aes(x = functionalGroup, y = dryMass),
    draw_quantiles = c(0.25, 0.5, 0.75))
```

```
## 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: The violin plot cannot adequately represent all the frequency points in order to produce a visualization plot. Given that it has a wider quartile range for some functional groupings, the boxplot is more helpful in this situation.

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

> Answer: Needles with the highest biomass/dry mass contribute the most to the litter.