Assignment 3: Data Exploration

Aishwarya Patankar

Spring 2025

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.

```r
#Installing the packages tidyverse, lubridate and here
#install.packages("tidyverse") : Line is hashed out while knitting file
#install.packages("lubridate")  Line is hashed out while knitting file
#install.packages("here")  Line is hashed out while knitting file

library(tidyverse)
library(lubridate)
library(here)

#Checking the working directory
getwd()
```

```
## [1] "/Users/aishwaryapatankar/Documents/Duke University/Spring2025/ENERGY872/EDA_Spring2025"
```

```
#Uploading 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: Neonicotinoids are a class of chemicals affecting the central nervous system of insects and are used as insecticides. However, these can impacts both harmful as well as helpful insects. They are harmful to pollinators and contaminate the soil, streams and wetlands. Their impact on pollinators especially can disrupt agricultural productivity, thus their study is of importance.

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 acts as a covering for the soil and prevents drying of the soil as well as soil erosion. The organic content of the material gets decomposed and becomes available for uptake by microorganisms, plants, shrubs and trees in the ecosystem thus continuing the carbon cycle. It also impacts waterflows, sediment transport and thereby impacts floodplains. All these factors contribute to our interest in studying this.

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. Ground and Elevated traps are used for sampling. 2. Mass reported are reported at the level of spatial resolution of single trap and temporal resolution of a collection event. 3. Material is classified in different functional groups such as leaves, twigs, needles, seeds, woody material etc.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
# Getting the dimensions of the dataset using the dim command
dim(Neonics)
```

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

```
# We see that there are 4623 observations (rows) and 30 variables (columns)
```

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…]

```
# Using the summary command and considering the "Effect" column.
# It is observed that the Effect column is the 19th column in the dataframe.
summary(Neonics[c(19)])
```

```
##               Effect
##  Population      :1803
##  Mortality       :1493
##  Behavior        : 360
##  Feeding behavior: 255
##  Reproduction    : 197
##  Development     : 136
##  (Other)         : 379
```

Answer:The effects of population, mortality, behaviours, feeding behaviour, reproduction and development are the most common effects of interest since we are observing the ecotoxicity of the insecticides. Insecticides directly impact the how many insects will survive, feed, reproduce, develop and thrive in the environment. It helps us estimate the efficacy of the insecticide as well as impact of the species that are present in the ecosystem

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…]

```
# Using the summary command and considering the "Species Common Name" column.
# It is observed that our column of interest is the 7th column in the dataframe.
summary(Neonics[c(7)])
```

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

Answer: The Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, Italian Honeybee are all pollinators and are essential to the process of pollination and thus creation of seeds, fruits and biodiversity of the ecosystem. Pollinators are of importance in the agrarian economy as they are responsible for agricultural food production.

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…]

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

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

Answer: We observe that there are multiple cells in the table with 'NR' written in them also multiple numbers in the column have a forward slash '/' written after the number. Indicating that the column is not numeric.

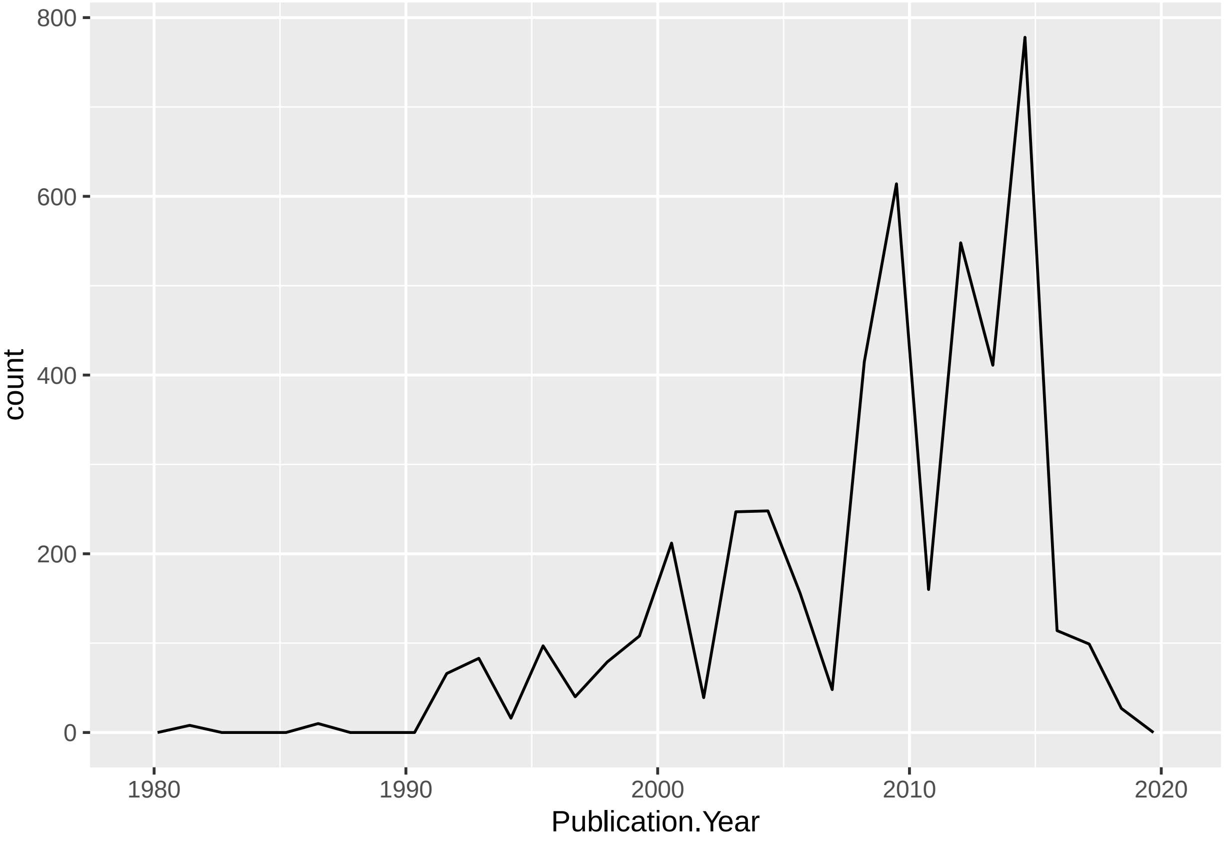
Explore your data graphically (Neonics)

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

```
# Creating a plot using ggplot and geom_freqpoly
ggplot(Neonics)+
  geom_freqpoly(aes(x=Publication.Year, bins = 30))
```

```
## Warning in geom_freqpoly(aes(x = Publication.Year, bins = 30)): Ignoring
## unknown aesthetics: bins
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
  ggtitle("Number of Studies")
```

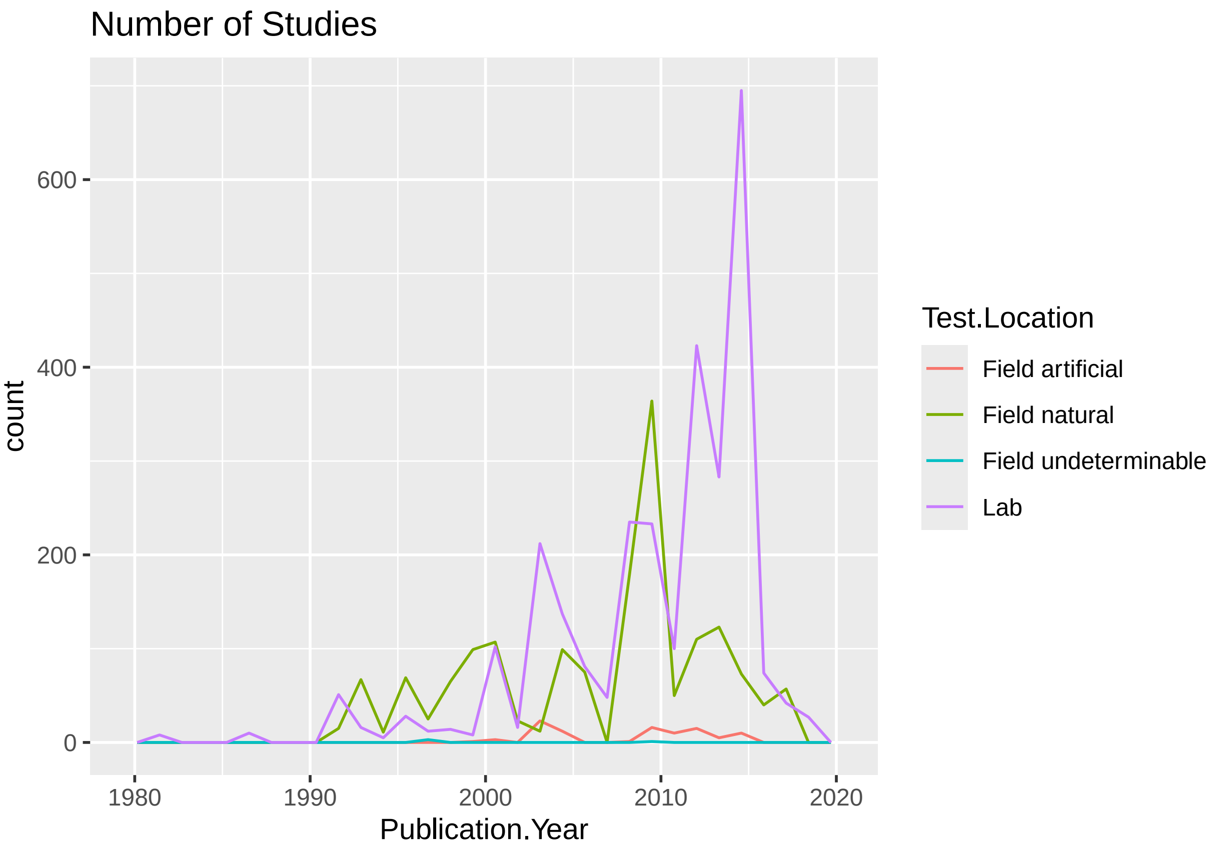
```
## $title
## [1] "Number of Studies"
##
## attr(,"class")
## [1] "labels"
```

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

```
# Creating a plot using ggplot and geom_freqpoly
ggplot(Neonics)+
  geom_freqpoly(aes(x=Publication.Year, colour = Test.Location, bins = 30 ))+
  ggtitle("Number of Studies")
```

```
## Warning in geom_freqpoly(aes(x = Publication.Year, colour = Test.Location, :
## Ignoring unknown aesthetics: bins
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

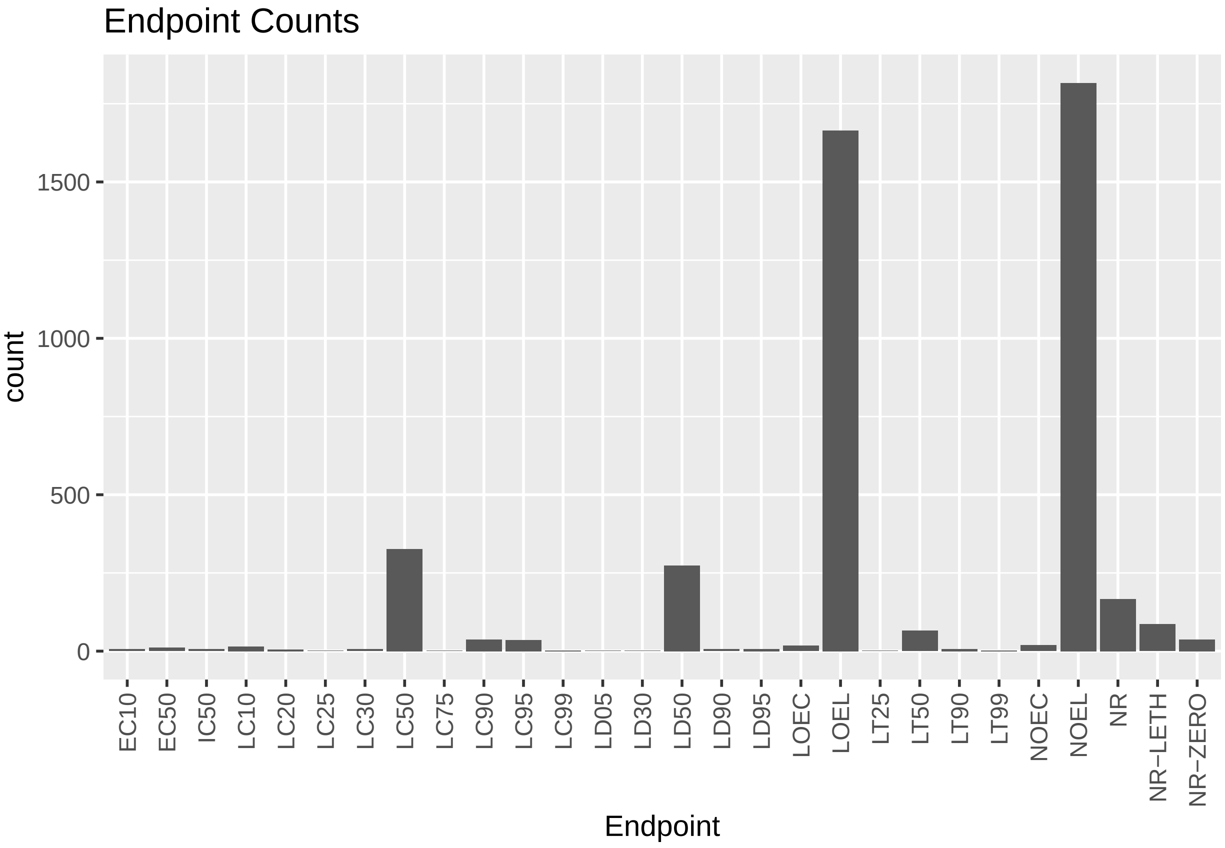
> Answer: The most common test locations is the Lab. Yes the test locations differ over time. We observe that upto 1990 there are very few Test and are mostly conducted in Lab. From Mid

1990's Upto 2000 we observe that Tests are mostly in Natural Field. From 2000 to 2008 Tests are in Lab. In 2009 there is a peak in natural Field tests. After 2010 a very high number of tests are conducted in the Lab.

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…]

```
ggplot(Neonics)+
  geom_bar(aes(x=Endpoint))+
  ggtitle("Endpoint Counts") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: The two most common end points are 'NOEL' and 'LOEL'. NOEL stands for 'No observable Effect Level' defined as Highest dose (concentration) producing effects not significantly different from responses of controls according to authors reported statistical test. "LOEL is 'Lowest observable effect level' defined as lowest dose (concentration) 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.

```r
#Getting the 'class' of the 'collectDate' column
class(Litter$collectDate)
```

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

```r
# Converting date from 'factor' to 'date format'
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")

#Reconfirming if new class of the variable
class(Litter$collectDate)
```

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

```r
#Getting Dates for Litter Sampling is August 2018
August <- Litter[Litter$collectDate >= "2018-08-01" & Litter$collectDate <= "2018-08-31", ]

unique(August$collectDate)
```

```
## [1] "2018-08-02" "2018-08-30"
```

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
# Using unique to find out the unique sampled plots
SampledPlots <- Litter$plotID
unique(SampledPlots)
```

```
##  [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
```

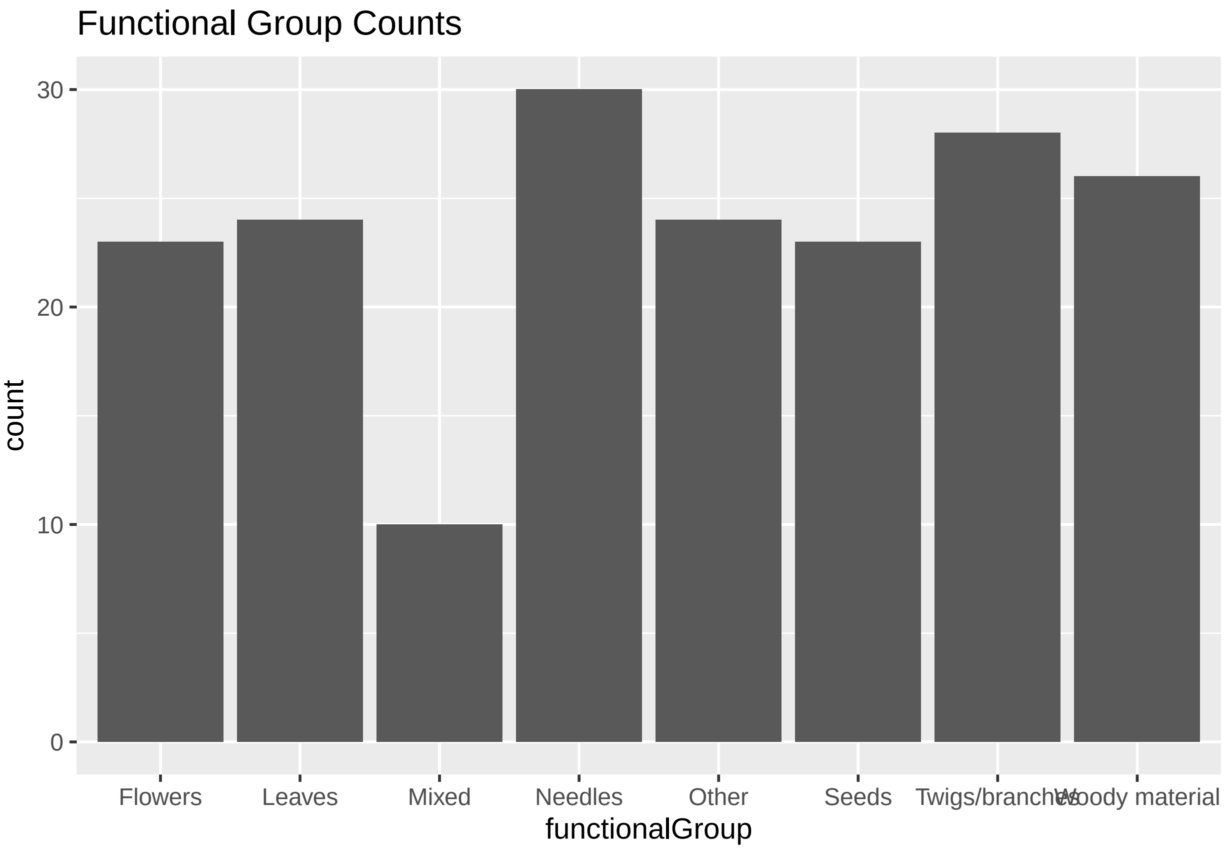
```r
# Using summary to find out summary of the Sampled Plots
summary(SampledPlots)
```

```
## 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: The 'unique option provides us a list of information about which are the unique plots that were sampled. Whereas summary provides us information of list of information about which are the unique plots that were sampled and number of samples collected at each unique plot Id.

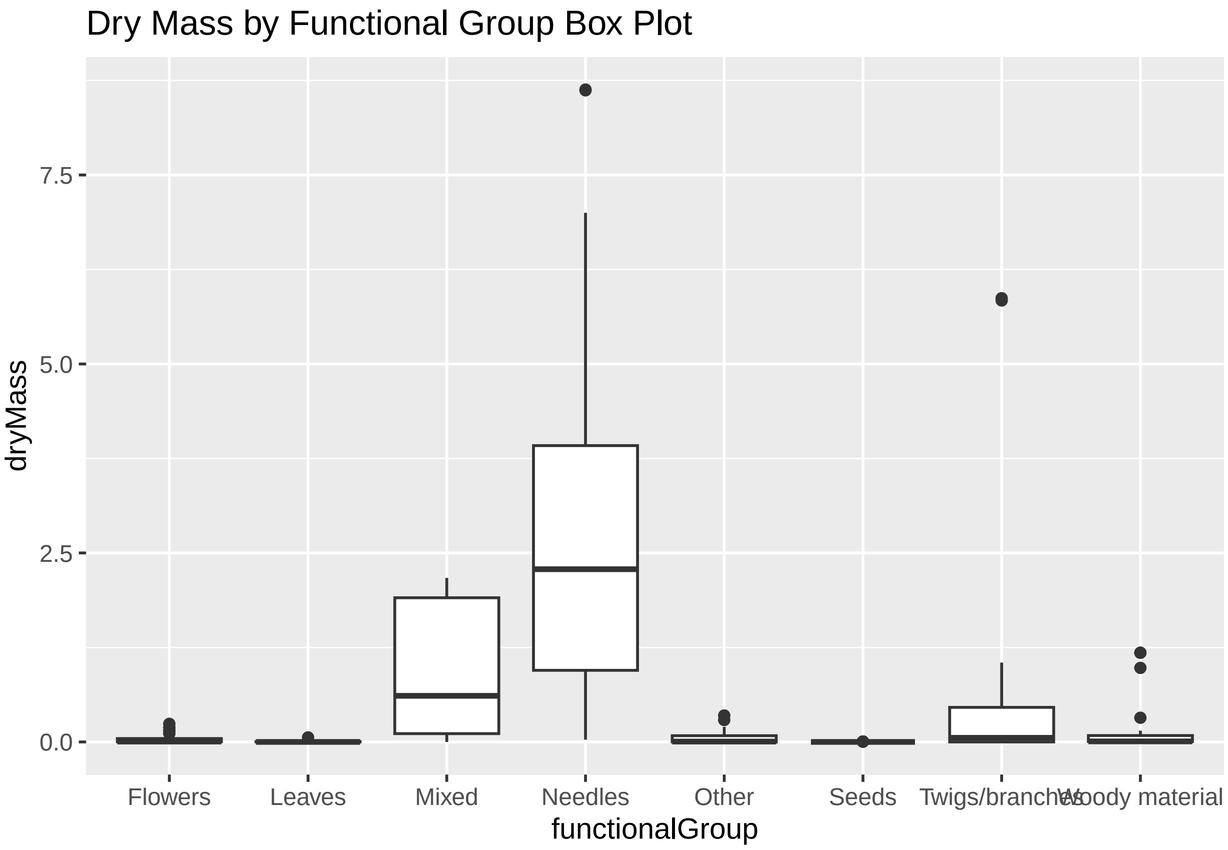
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
ggplot(Litter)+
  geom_bar(aes(x=functionalGroup))+
  ggtitle("Functional Group Counts")
```

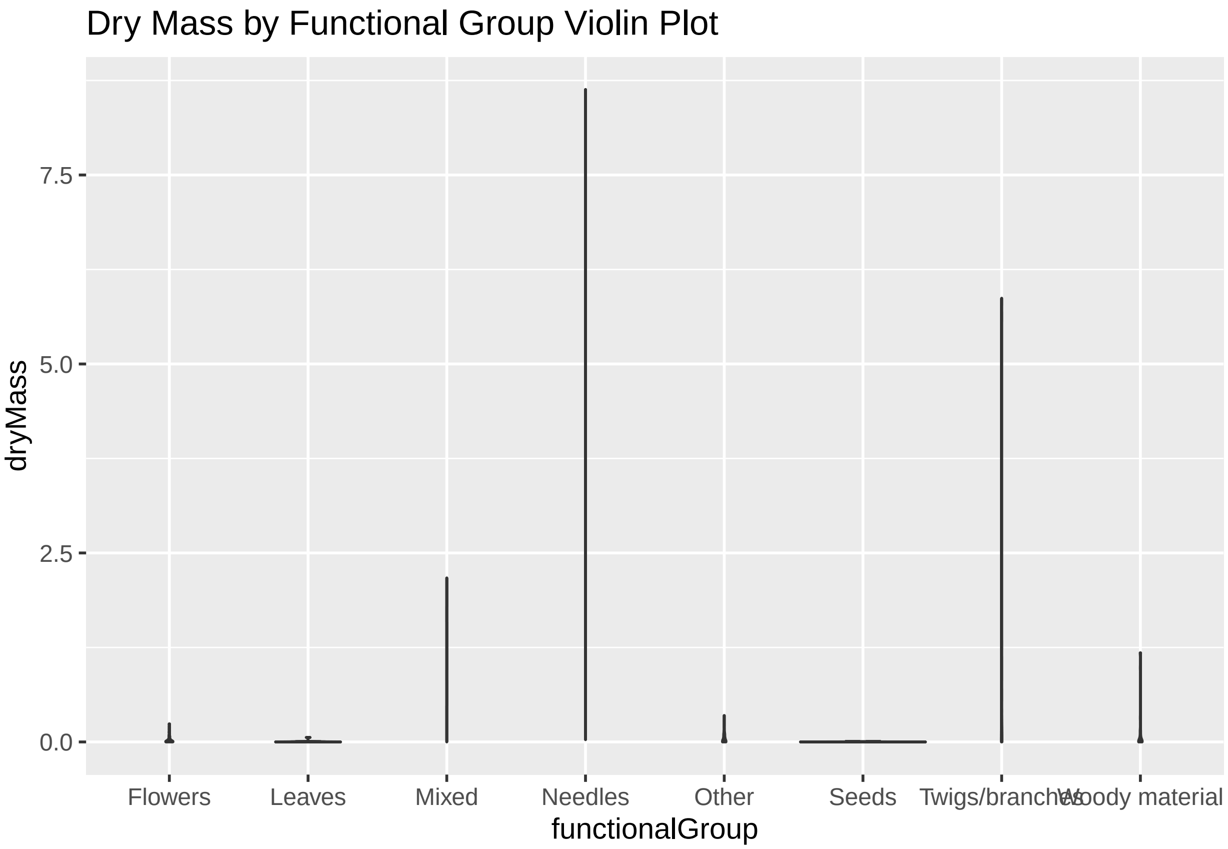


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#Creating a Boxplot of Dry Mass by Functional Group
ggplot(Litter)+
  geom_boxplot(aes(y=dryMass,x=functionalGroup))+
  ggtitle("Dry Mass by Functional Group Box Plot")
```



```
# Creating a Violin plot of Dry Mass by Functional Group
ggplot(Litter)+
  geom_violin(aes(y=dryMass,x=functionalGroup),
  draw_quantiles = c(0.25, 0.5, 0.75)) +
  ggtitle("Dry Mass by Functional Group Violin Plot")
```



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

> Answer: The boxplot is a better visualization as it helps us to see the median and the quartile ranges even in a spread out distribution. Whereas, in a violin plot we are unable to see the distribution since the values are spread out.

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

> Answer: Needles form the highest biomass at the sites.