Assignment 3: Data Exploration

Aadya Shukla

Spring 2026

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. [NEW] 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 Canvas.
8. Initial here to acknowledge that you did not use AI in completing this assignment, except where expressly allowed: _____

TIP: If your code extends past the page when knit, tidy your code by manually inserting line breaks in your code chunks.

TIP: If your code fails to knit, check: * That no `install.packages()` or `View()` commands exist in your code. * That you are not displaying the entire contents of a large dataframe in your code.

Set up your R session

1. Load necessary packages (tidyverse, here), check your current working directory and import 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: * Use the `here()` package in specifying the paths to your datasets * Include the appropriate subcommand to read in character based columns as factors

```
library(tidyverse)
library(here)
#checking working directory
here()
```

```
## [1] "/home/guest/872L/EDE_Spring2026"
```

```
#Adding raw data
Neonics <- read.csv(
  file = here('Data','Raw','ECOTOX_Neonicotinoids_Insects_raw.csv'),
  stringsAsFactors = T
)

Litter <- read.csv(
  file = here('Data','Raw','NEON_NIWO_Litter_massdata_2018-08_raw.csv'),
  stringsAsFactors = T
)
```

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. (AI is allowed here, but put answers in your own words.)

   Answer: It is important to understand the effect of neonicotinoids as the wide usage could have implications on key biological systems as well as food and human health. Neonicotinoids must be evaluated for their negative impacts on pollinators and insects that predate on pests as their survival determines crop health and vegetation diversity. Any toxins present in neonicotinoids could also lead to biomagnification, exposing multiple species down the food chain to harmful chemicals. Additionally, any leeching of toxins into our soils and water systems would cause widespread contamination. Thus, ecotoxicology of neonicotinoids is necessary to prevent harm to non-target species and accumulating impacts to diverse ecosystems.

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. (AI is allowed here, but put answers in your own words.)

   Answer: Litter and woody debris play key roles in several ecological cycles. As debris falls to the ground, it provides shelter to various organisms as well as acting as feeding and breeding grounds for them. The debris is also an important element in carbon sequestration and the nitrogen and phosphorus cycles. Litter can also often protect the soil from intense atmospheric conditions and prevent soil erosion. Thus, studying litter and woody debris can help explain some natural mechanisms necessary to keep our forests healthy and long-term changes in, and impacts to, forest health.

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. Different physical tools are employed based on litter dimensions, including elevated PVC traps and ground traps. 2. Overall sampling locations are selected at random with variations in number of plots and trap placement according to vegetation type. 3. Ground traps are sampled once every year while elevated traps are sampled according to vegetation type.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#view(Neonics)
class(Neonics)
```

```
## [1] "data.frame"
```

```
str(Neonics)
```

```
## 'data.frame':    4623 obs. of  30 variables:
##  $ CAS.Number                       : int  58842209 58842209 58842209 58842209 58842209 58842209 5884
##  $ Chemical.Name                    : Factor w/ 9 levels "(1E)-N-[(6-Chloro-3-pyridinyl)methyl]-N-ethy
##  $ Chemical.Grade                   : Factor w/ 9 levels "Analytical grade",..: 9 9 9 9 9 9 9 9 9 9 .
##  $ Chemical.Analysis.Method         : Factor w/ 5 levels "Measured","Not coded",..: 4 4 4 4 4 4 4 4 4
##  $ Chemical.Purity                  : Factor w/ 80 levels ">=98",">=99.0",..: 69 69 50 50 50 50 50 50
##  $ Species.Scientific.Name          : Factor w/ 398 levels "Acalolepta vastator",..: 69 69 248 248 24
##  $ Species.Common.Name              : Factor w/ 303 levels "Alfalfa Leafcutter Bee",..: 74 74 142 142
##  $ Species.Group                    : Factor w/ 4 levels "Insects/Spiders",..: 1 1 1 1 1 1 1 1 1 1 ..
##  $ Organism.Lifestage               : Factor w/ 20 levels "Adult","Cocoon",..: 1 1 19 19 19 1 19 1 1
##  $ Organism.Age                     : Factor w/ 39 levels "<=24","<=48",..: 39 39 39 39 39 36 39 36 3
##  $ Organism.Age.Units               : Factor w/ 11 levels "Day(s)","Days post-emergence",..: 9 9 4 4
##  $ Exposure.Type                    : Factor w/ 24 levels "Choice","Dermal",..: 23 23 11 11 11 11 11
##  $ Media.Type                       : Factor w/ 10 levels "Agar","Artificial soil",..: 7 7 3 3 3 3 3
##  $ Test.Location                    : Factor w/ 4 levels "Field artificial",..: 4 4 4 4 4 4 4 4 4 4 .
##  $ Number.of.Doses                  : Factor w/ 30 levels "’ 4-5","’ 4-7",..: 30 30 18 18 18 18 18 18
##  $ Conc.1.Type..Author.             : Factor w/ 3 levels "Active ingredient",..: 1 1 1 1 1 1 1 1 1
##  $ Conc.1..Author.                  : Factor w/ 1006 levels "<0.0004","<0.025",..: 639 510 813 622 44
##  $ Conc.1.Units..Author.            : Factor w/ 148 levels "%","% v/v","% w/v",..: 132 132 91 91 91 9
##  $ Effect                           : Factor w/ 19 levels "Accumulation",..: 16 16 16 16 16 16 16 16
##  $ Effect.Measurement               : Factor w/ 155 levels "Abundance","Accuracy of learned task, per
##  $ Endpoint                         : Factor w/ 28 levels "EC10","EC50",..: 15 15 8 8 8 8 8 8 8 8 ...
##  $ Response.Site                    : Factor w/ 19 levels "Abdomen","Brain",..: 14 14 14 14 14 14 14
##  $ Observed.Duration..Days.         : Factor w/ 361 levels "<.0002","<.0021",..: 145 145 145 145 145
##  $ Observed.Duration.Units..Days.   : Factor w/ 17 levels "Day(s)","Day(s) post-emergence",..: 1 1 1
##  $ Author                           : Factor w/ 433 levels "Abbott,V.A., J.L. Nadeau, H.A. Higo, and
##  $ Reference.Number                 : int  107388 107388 103312 103312 103312 103312 103312 103312 10
##  $ Title                            : Factor w/ 458 levels "A Common Pesticide Decreases Foraging Suc
##  $ Source                           : Factor w/ 456 levels "Acta Hortic.1094:451-456",..: 295 295 296
##  $ Publication.Year                 : int  1982 1982 1986 1986 1986 1986 1986 1986 1986 1986 ...
##  $ Summary.of.Additional.Parameters: Factor w/ 943 levels "Purity: \xca NC - NC | Organism Age: \xca
```

```r
dim(Neonics)
```

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

```r
#There are 4,623 rows and 30 columns
```

6. Using the summary function on the "Effect" column, determine the most common effects that are studied. [Tip: The sort() command is useful for listing the values in order of magnitude…]

```r
#Using summary and sort to identify the most common to least common effects
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
```

```r
sort(
  summary(Neonics$Effect),
  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
```

Question: Which two effects stand out as the most studied? Can you guess why these effects might specifically be of interest? > Answer: Population and mortality are the most commonly studied effects of neonicotinoids. These are important metrics to evaluate as researchers want to understand how effective the insecticide might be at removing target organisms while also identifying which non-target organisms are being affected and to what extent. If too many of non-target organisms are being impacted, it could affect the entire ecosystem.

7. Using the summary function, determine the six most commonly studied species in the dataset (common name).[TIP: Explore the help on the summary() function, in particular the maxsum argument…]

```r
#help("summary")

summary(Neonics$Species.Common.Name,maxsum = 6)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee 
##                   667                   285                   183 
##    Carniolan Honey Bee            Bumble Bee               (Other) 
##                   152                   140                  3196
```

Question: What do these species have in common? Why might they be of interest over other insects? > Answer: The six most commonly studied species include bees and wasps. These species are all beneficial insects that provide pollinating and pest control services to our crops. If these species are frequently targeted by insecticides then plant productivity and diversity will immensely reduce. Thus, it is important to understand to what extent neonicotinoids are affecting them.

8. The `Conc.1..Author` column, which lists the concentration of the neonicotinoid dose, should include numeric values. 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…]

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

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

```r
#view(Neonics)
```

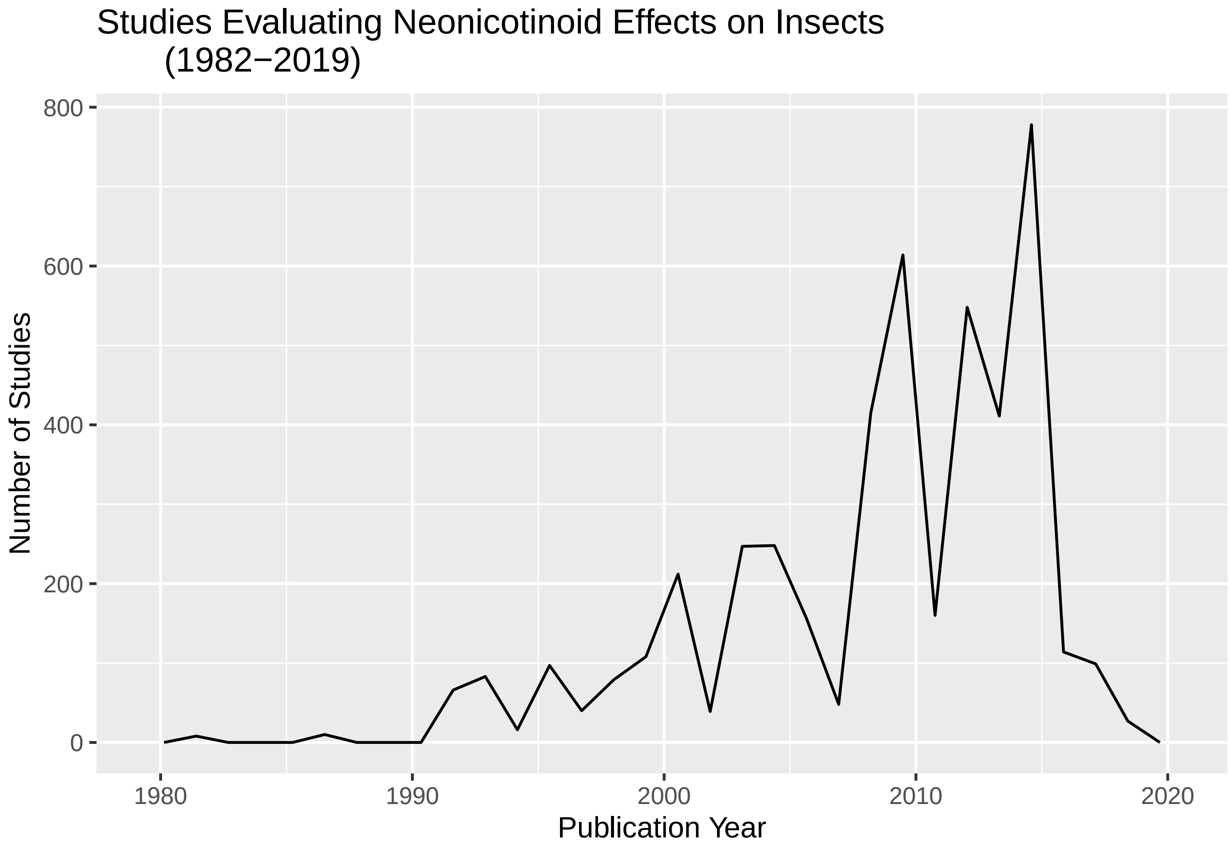
> Answer: The column 'Conc.1..Author' has a factor class. It is not numeric because several entries are not solely numbers. Some include '/' next to numbers to represent a ratio and some entries are labeled as 'NR', not the same as 'NA' which is still allowed for numeric class.

Explore your data graphically (Neonics)

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

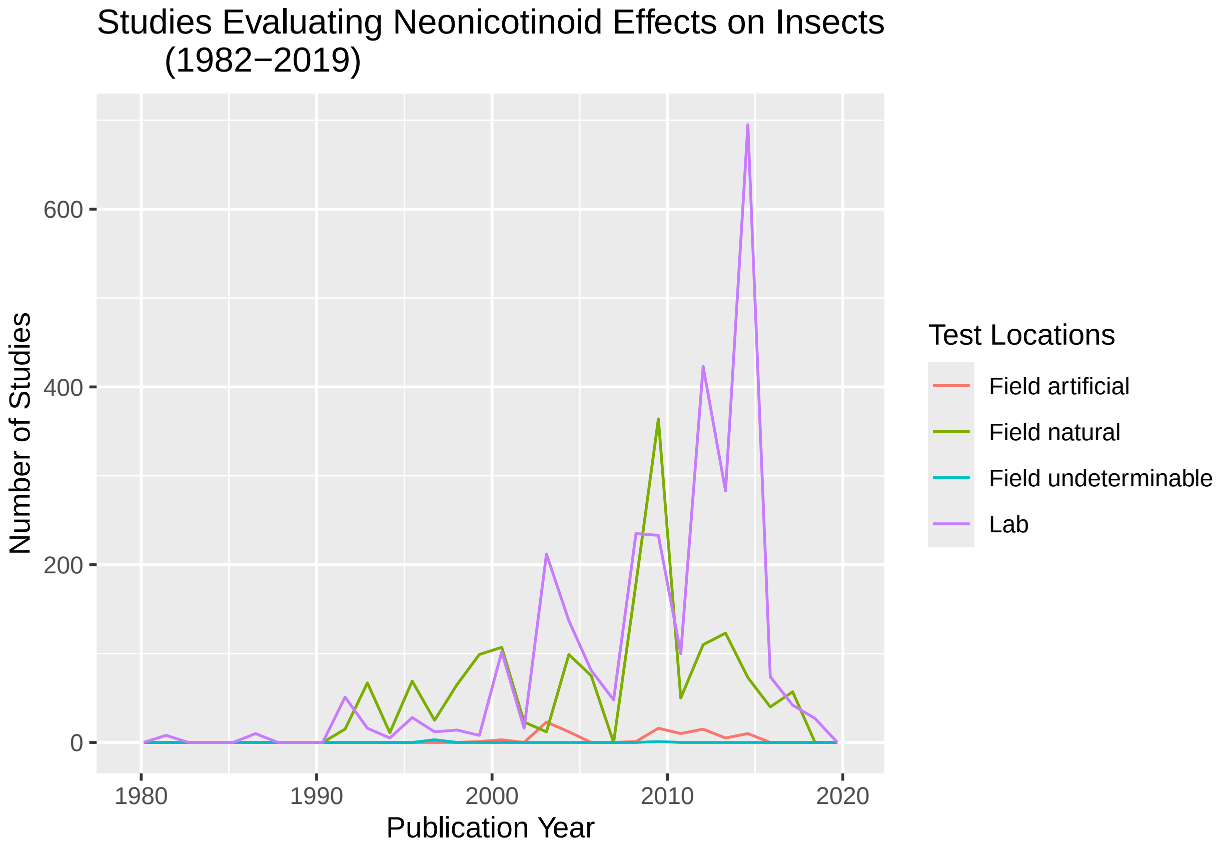
```r
library(ggplot2)

# Generating a frequency plot with x and y axis labels and a title using labs()
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 30) +
  labs(x = "Publication Year", y = "Number of Studies",
       title = "Studies Evaluating Neonicotinoid Effects on Insects
       (1982-2019)"
       )
```



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

```r
# Renewed code with test locations as an added variable; changed legend title
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 30) +
  labs(x = "Publication Year", y = "Number of Studies",
       title = "Studies Evaluating Neonicotinoid Effects on Insects
       (1982-2019)",
       color = "Test Locations"
       )
```

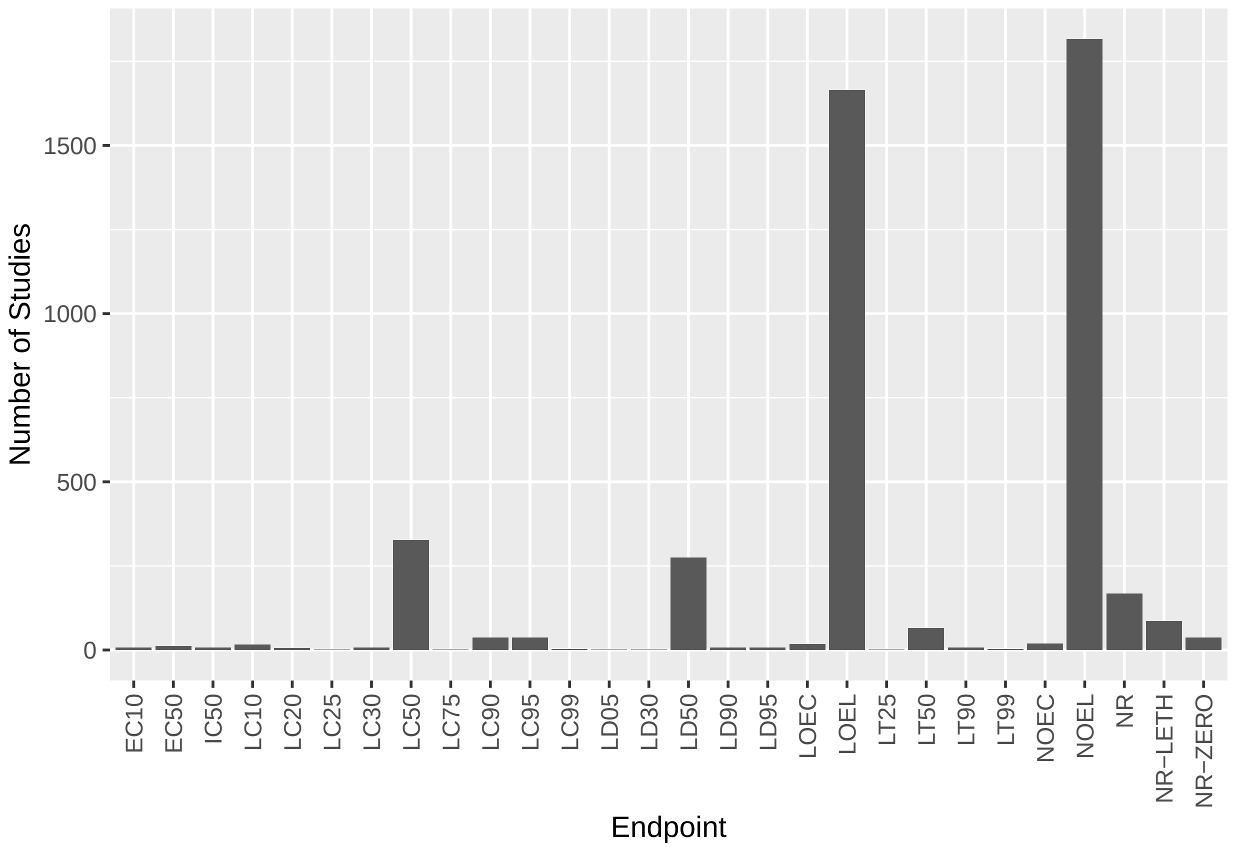


nterpret this graph. What are the most common test locations, and do they differ over time? > Answer: The most common test locations seem to be labs and natural fields. Over time the usage of labs in testing neonicotinoids has increased a lot more than testing in natural fields. The usage of natural fields also decreased sharply before 2010 and has stayed comparitively extremely low.

11. Create a bar graph of Endpoint counts.

[**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(data = Neonics, aes(x = Endpoint)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  labs( y = "Number of Studies")
```



What are the two most common end points, and how are they defined? Consult the ECOTOX_CodeAppendix (p.721) for more information. > Answer: The two most common end points are NOEL and LOEL. NOEL stands for ecotox levels to no observable effects, meaning the highest concentration levels do not behave differently compared to the controls used by the author. LOEL similarly stands for low observable effects, meaning the lowest concentration levels behaved significantly differently compared to the controls used.

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
class(Litter$collectDate) #class is factor, not date
```

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

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

class(Litter$collectDate) #class is now date
```

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

```r
unique(Litter$collectDate) #litter was sampled on 08-02 and 08-30 in 2018
```

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

13. Using the `unique` function, list the different `plotIDs` sampled at Niwot Ridge.

```r
unique(Litter$plotID) # there are 12 unique plot IDs
```

```
##  [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
```

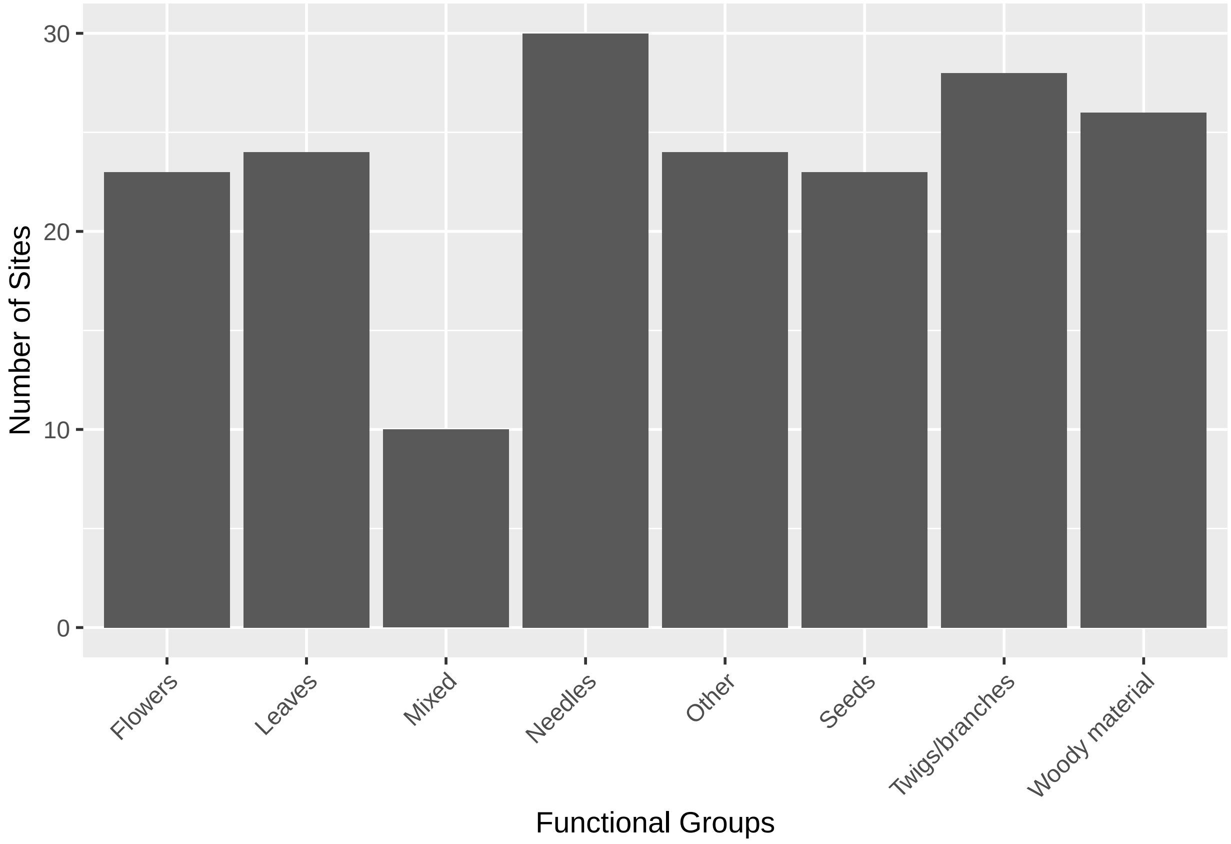
```r
summary(Litter$plotID)
```

```
## 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
```

How is the information obtained from `unique` different from that obtained from `summary`? > Answer: Using summary for the plotID column shows us the frequency at which data was collected from each plot. However, if we only wanted to know exactly how many unique plots exist, the unique function makes it easy for us by identifying that there are 12 distinct levels in the column. This may not make too much of a difference in small datasets where we could count the results from the summary. However, in larger datasets that have several more distinct values, using the unique function would simplify the process of identifying how many distinct plots exist.

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.

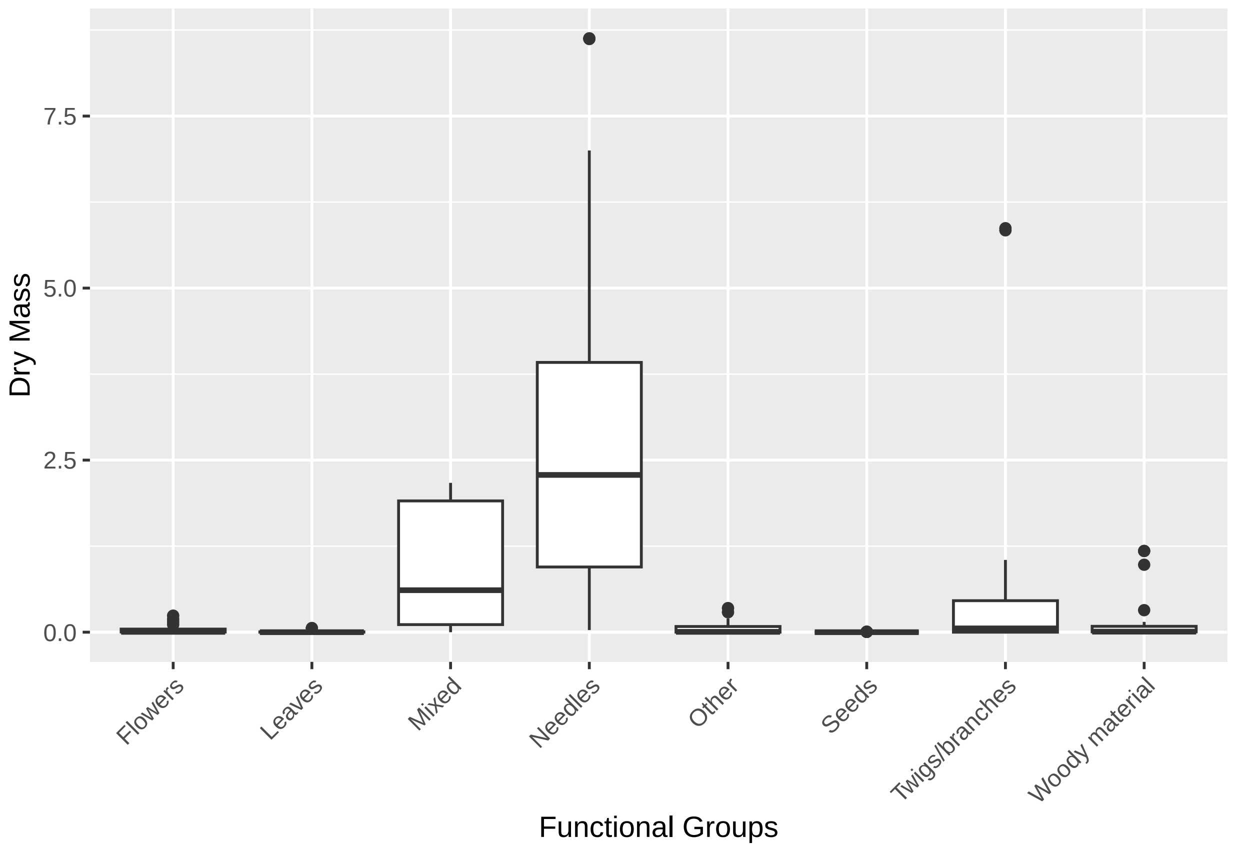
```r
ggplot(data = Litter, aes(x = functionalGroup)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  +
  labs(x = "Functional Groups", y = "Number of Sites")
```



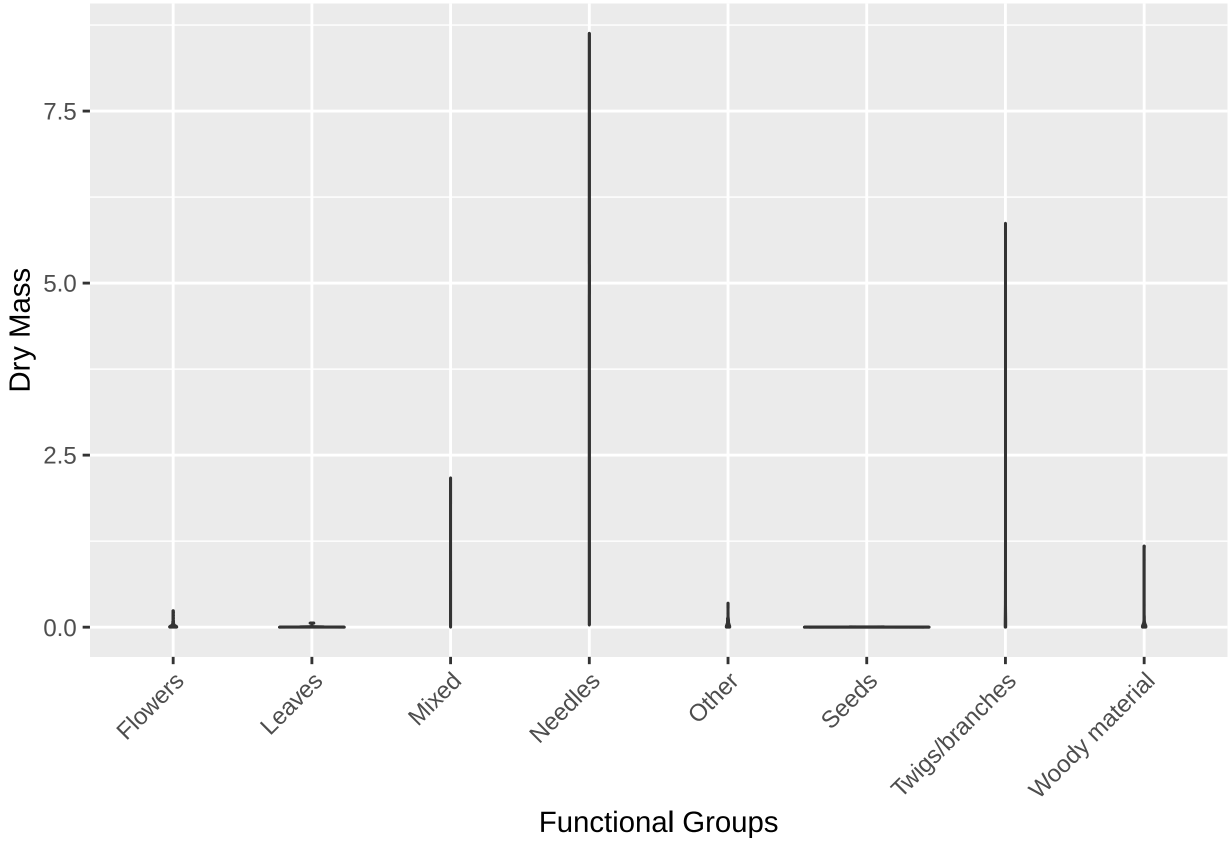
15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of `dryMass` by `functionalGroup`.

```
#Boxplot

ggplot(Litter) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass,
                   group = cut_width(functionalGroup, 1))) +
  labs(x = "Functional Groups", y = "Dry Mass")
```



```
#Violin plot
ggplot(Litter) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75)) +
  labs(x = "Functional Groups", y = "Dry Mass")
```



Why is the boxplot a more effective visualization option than the violin plot in this case? > Answer: We saw from the bar graph that the litter groups are fairly distributed across all sites. A violin plot represents the number of values as proportional to the width of the plot. Since the number of values is similar, the width of the violin plots is not telling us much. Additionally, the violin plots are exaggerating the overall trend of biomass distribution at the sites. The boxplot is effective because it shows us the distribution of biomass with a clear view of the median values and the outliers.

What type(s) of litter tend to have the highest biomass at these sites? > Answer: Needles seem to make up the highest biomass at these sites, followed by a mixed variety of functional groups.