Assignment 3: Data Exploration

Humayra Rahman

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: ___HR_

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

```
#Import packages
library(tidyverse); library(here)

Neonics <- read_csv(here("Data", "Raw", "ECOTOX_Neonicotinoids_Insects_raw.csv"))

Litter  <- read_csv(here("Data", "Raw", "NEON_NIWO_Litter_massdata_2018-08_raw.csv"))
```

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: We care about neonicotinoids and insects because these pesticides are designed to affect insect nervous systems, so they can also harm beneficial insects like pollinators and natural pest controllers, not just target pests. Since they can move through plants, soil, and water, insects may be exposed in the real world through pollen and nectar or contaminated habitats. The impacts can be obvious (death) or subtle (changes in behavior or reproduction), and either way they can ripple through ecosystems and even affect crop productivity and biodiversity.

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:We study forest litter and woody debris because it's how forests "recycle" themselves: fallen leaves, needles, and branches decompose and return carbon and nutrients to the soil, shaping soil fertility and long term carbon storage. Tracking how much falls over time also tells us about forest productivity and how the ecosystem is responding to changes like drought, pests, storms, or wildfire risk from built up fuels.

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.NEON uses two traps: raised mesh baskets catch leaves and short twigs, and marked ground strips catch long thin sticks. 2.Traps are set up in forest plots with trees over about 2 m tall, with one raised and one ground trap pair for each standard area of the plot. 3. Ground traps are usually collected once a year, but the raised baskets are emptied more often, especially during fall leaf drop.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
dim(Neonics) # shows the number of rows and columns in the neonics dataset
```

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

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
eff_tab <- table(Neonics$Effect) # shows the how many times each effect appears by counting them

sort(eff_tab, decreasing = TRUE) # from most effects --> least common effects
```

```
## 
##       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
```

```r
summary(Neonics$Effect) #summarizes the effect column
```

```
##    Length     Class      Mode 
##      4623 character character
```

Question: Which two effects stand out as the most studied? Can you guess why these effects might specifically be of interest? > Answer:The two effects that stand out are Population and Mortality. They're probably studied the most because they're the clearest, easiest ways to see harm: mortality shows whether the pesticide kills insects, and population effects show whether it could lower insect numbers over time, which is a big ecological and agriculture concern.

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
Neonics$`Species Common Name` <- as.factor(Neonics$`Species Common Name`) # turned character to factor

summary(Neonics$`Species Common Name`, maxsum = 6) # shows the six most commonly studied speices
```

```
##            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:They are mostly helpful insects, especially bees and bumblebees that pollinate plants. Some wasps are also beneficial because they help control pests. Researchers focus on these insects because they support crops and wild plants, and they can be exposed to neonicotinoids when they visit flowers, so any harm to them can affect pollination and food production.

8. The `Conc.1..Author` column, which lists the concentration of the neonicitoid 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...]

```
class(Neonics$`Conc 1 (Author)`) # checks what kind of data the column is
```

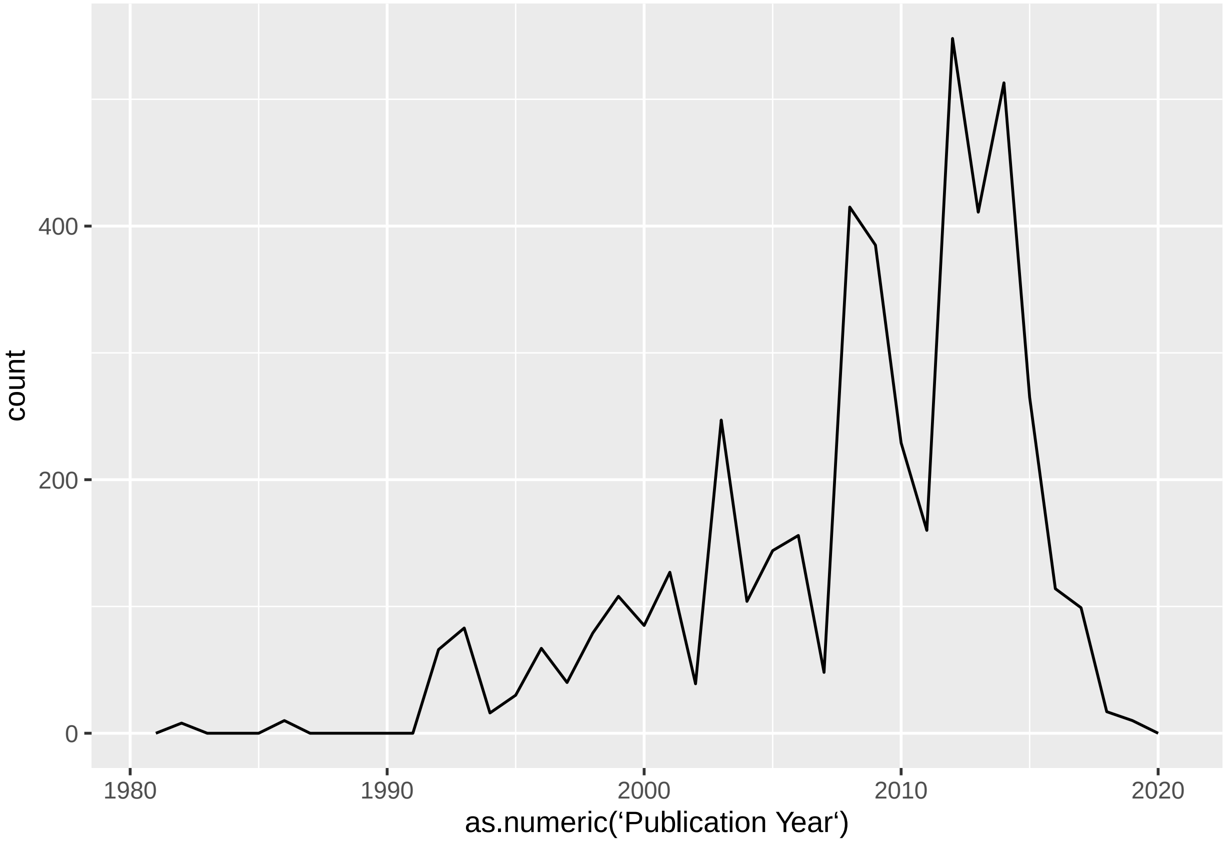
```
## [1] "character"
```

Answer:It is not numeric because many entries are not just numbers. Some include units or extra text like mg/L, symbols like < or >, ranges like 0.1 to 1.0, or codes and blanks like NA, ND, or NR. When that happens, R treats the whole column as text instead of numbers.

Explore your data graphically (Neonics)

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

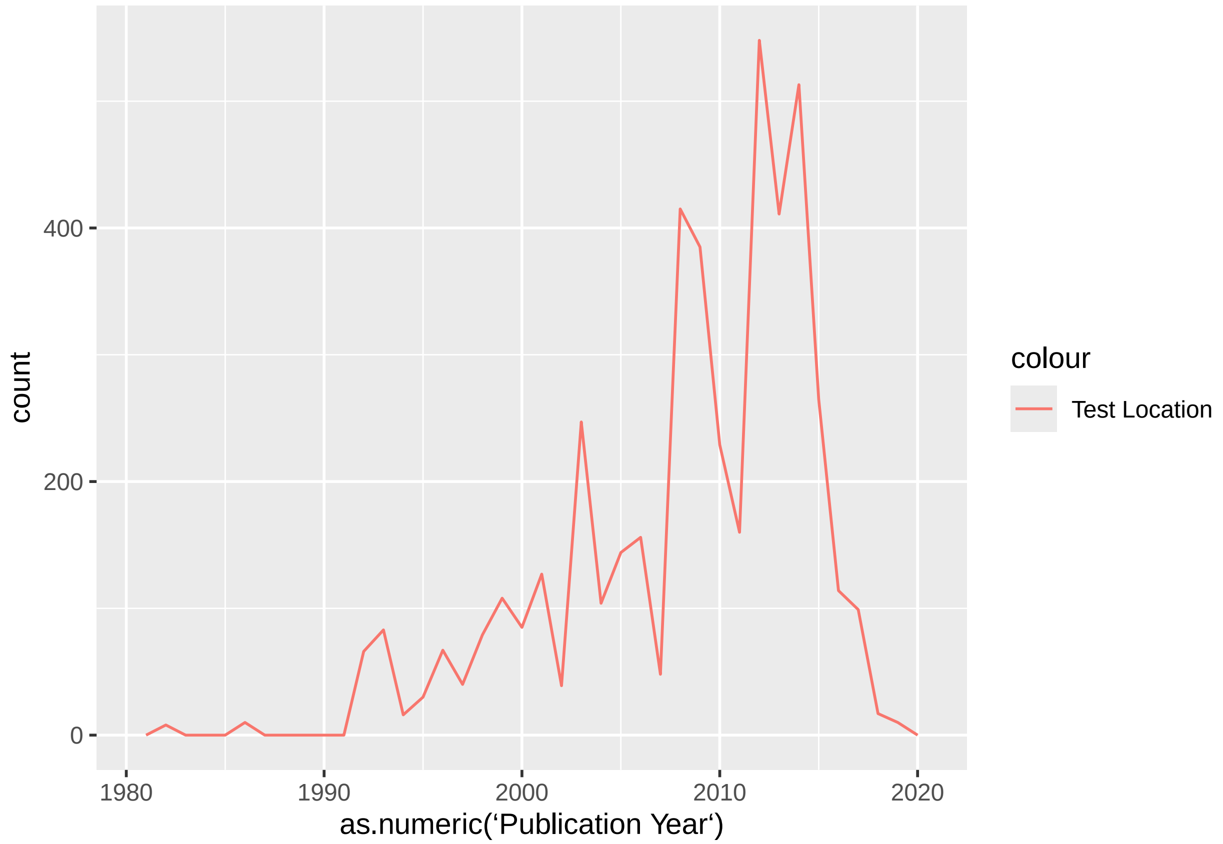
```
library(tidyverse) # loads ggplot and other plots

ggplot(data = Neonics, aes(x = as.numeric(`Publication Year`))) +
  geom_freqpoly(binwidth = 1)
```



```
# put publication year on the x axis + counts studies per year to draw a line and puts them in a bin of
```

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

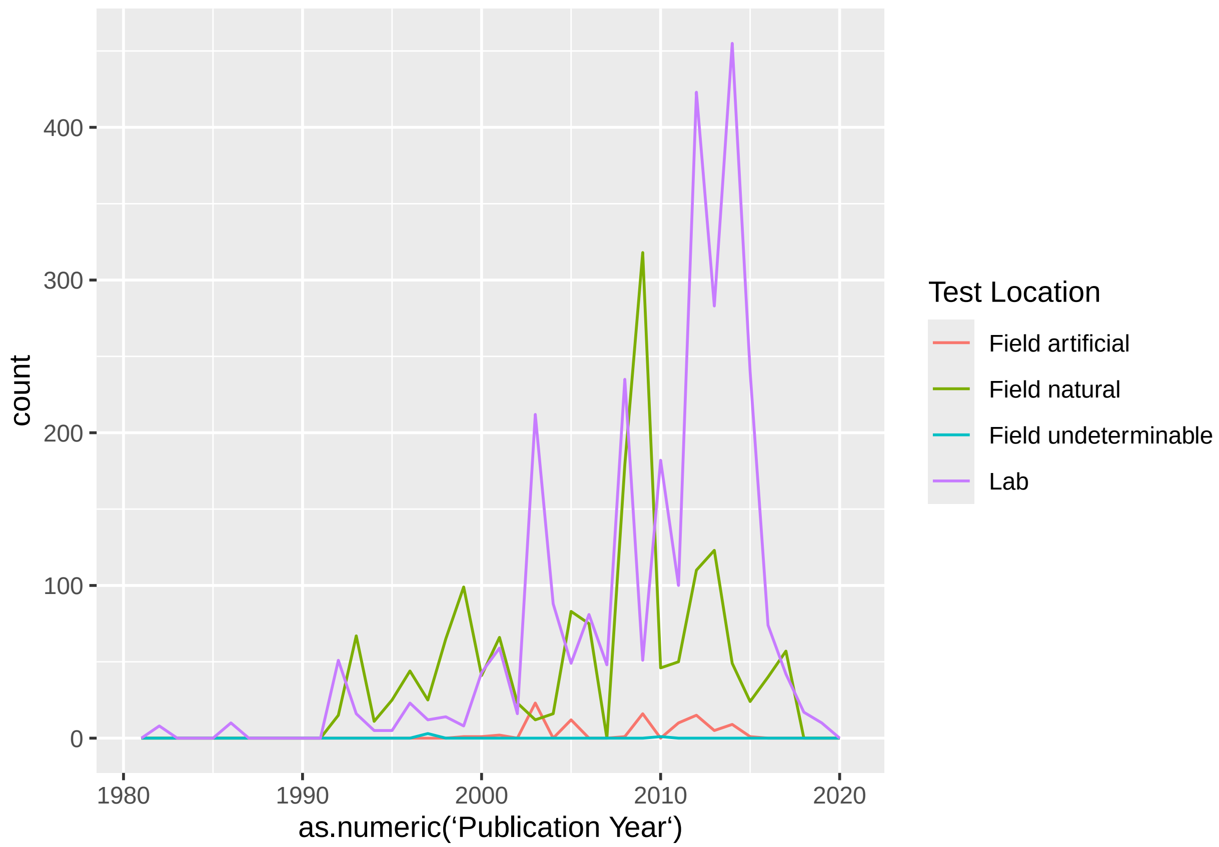
```r
ggplot(data = Neonics, aes(x = as.numeric(`Publication Year`), color = 'Test Location')) +
  geom_freqpoly(binwidth = 1) # added 1 colour to the graph
```



```r
table(Neonics$`Test Location`)
```

```
##
##     Field artificial        Field natural Field undeterminable
##                   96                 1663                    4
##                  Lab
##                 2860
```

```r
ggplot(Neonics, aes(x = as.numeric(`Publication Year`),
                    color = `Test Location`,
                    group = `Test Location`)) +
  geom_freqpoly(binwidth = 1) # added different colours to different test locations
```

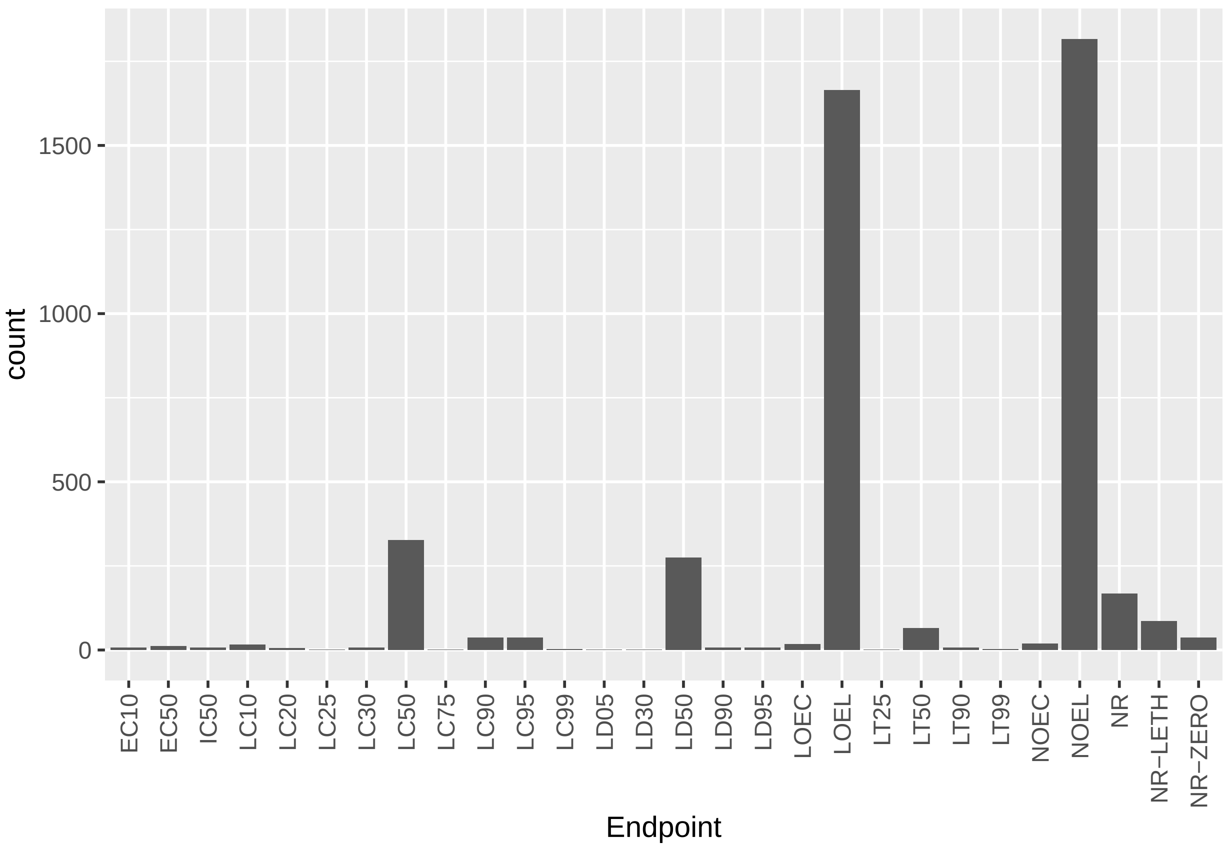


Interpret this graph. What are the most common test locations, and do they differ over time? > Answer:Lab is the most common test location by far, and it peaks strongly in the late 2000s to early 2010s. Field natural is the second most common, with smaller peaks spread across the 1990s to 2010s. Field artificial and field undeterminable are rare and stay 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...]

```r
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar() + # set the endpoint on x axis + made a bar chart
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



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 endpoints are LOEL and NOEL.LOEL is the lowest observable effect level, meaning the lowest dose or concentration that causes an effect that is significantly different from the control.NOEL is the no observable effect level, meaning the highest dose or concentration that does not cause an effect that is significantly different from the control.

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.

```
library(lubridate) # to work with dates

class(Litter$collectDate) # to check the data type
```

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

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

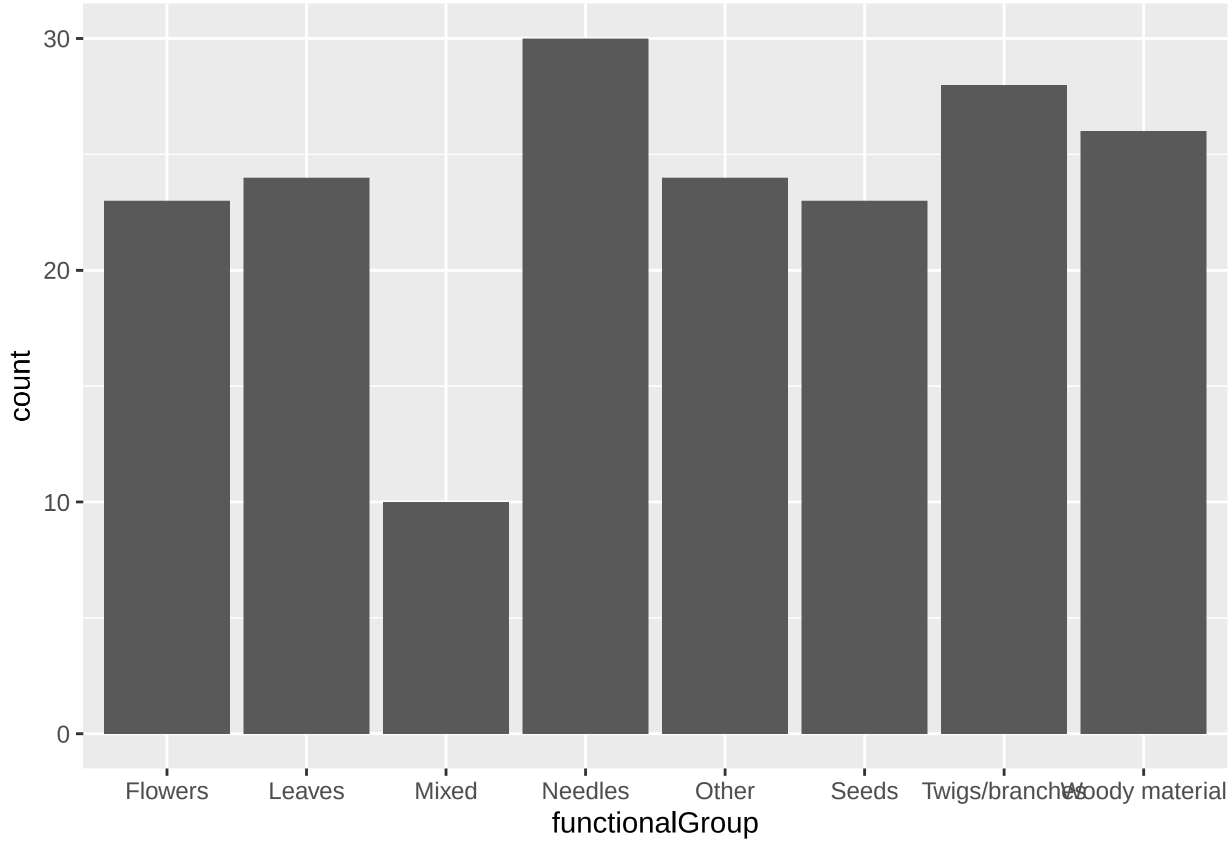
```r
sort(unique(Litter$plotID)) # sort all unique plot ID
```

```
##  [1] "NIWO_040" "NIWO_041" "NIWO_046" "NIWO_047" "NIWO_051" "NIWO_057"
##  [7] "NIWO_058" "NIWO_061" "NIWO_062" "NIWO_063" "NIWO_064" "NIWO_067"
```

How is the information obtained from `unique` different from that obtained from `summary`? > Answer:unique shows the exact different values that appear. summary gives a quick overview, like how often values show up or basic stats if it is numbers, instead of listing everything.

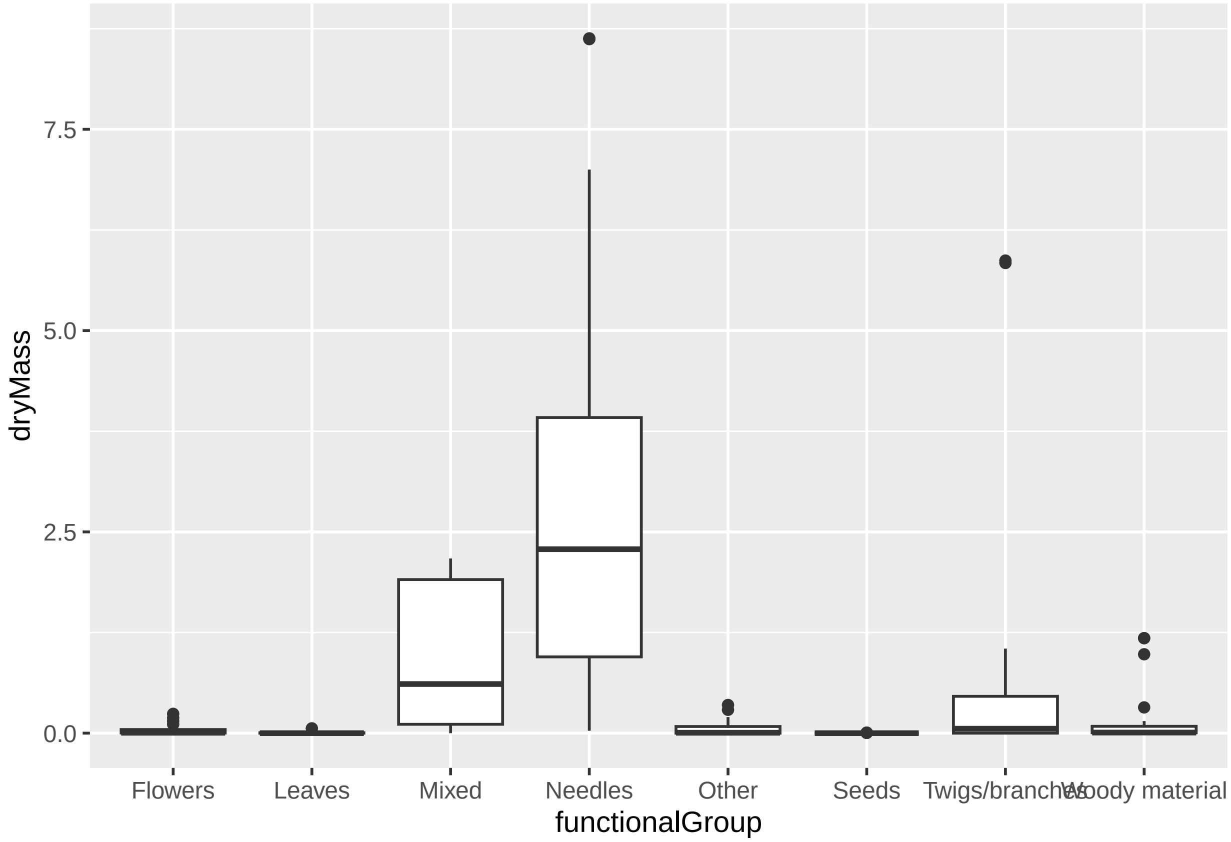
14. Create a bar graph of `functionalGroup` counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar() # puts the functionalgroup on x axis + makes a bar chart
```

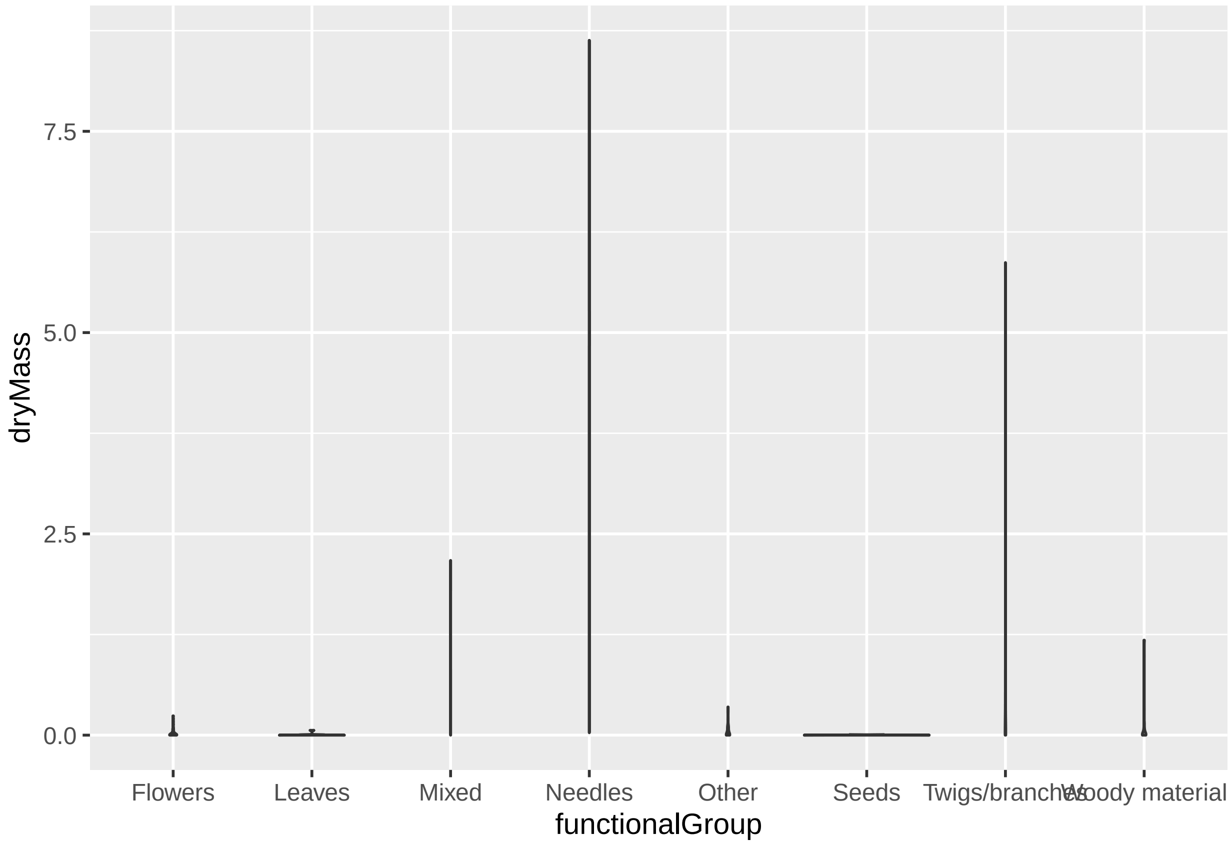


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of `dryMass` by `functionalGroup`.

```r
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_boxplot() # comapres drymass across functionalgroups + makes a bar chart
```



```
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_violin() # makes a violin plot
```



Why is the boxplot a more effective visualization option than the violin plot in this case? > Answer:The violin plot does not work well here because most dryMass values are near zero and the data are very skewed, so the violins collapse into thin shapes and are hard to read. The boxplot is clearer because it shows the median, spread, and outliers even with lots of near zero values.

What type(s) of litter tend to have the highest biomass at these sites? > Answer:Needles have the highest biomass overall. Mixed litter is next highest, and twigs and branches also show occasional high values but are usually lower than needles.