Assignment 3: Data Exploration

Lexi Nelson

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: AN

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

```r
#setup R session and load necessary packages
library(tidyverse) #load tidyverse
library(here) #load here as we worked with in Lab 3

here() #confirm where here command will point
```

```
## [1] "/home/guest/ENV872/EDE_Spring2026/Assignments"
```

```r
getwd() #check working directory
```

```
## [1] "/home/guest/ENV872/EDE_Spring2026/Assignments"
```

```r
#I coped the two data files from the EDE_Spring2026 raw data folder into a Data
#folder under my Assignments folder. I did this so that R could access the data
#in my working directory using "here".

#read in Neonicotinoids data set
Neonics <- read.csv(
  file = here('Data','ECOTOX_Neonicotinoids_Insects_raw.csv'),
  stringsAsFactors = T
)
#stringsasFactors is the command to read in character based columns as factors

#read in NEON data set
Litter <- read.csv(
  file = here('Data','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: Neonicotinioids are designed to protect insects by killing plants. We
are interested in the ecotoxicology of neonicotinoids on insects because they
are directly impacted. We would like to understand the mechanisms by which the
toxins act on insects. Additionally, insects underpin the entire food web, so
there could be cascading effects of the neonicotinoid use.


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 contain nutrients such as CO2 and others. These
molecules can decompose and release nutrients back to the atmosphere and soil.
They also serve as a habitat for some creatures and prevent erosion. Studying
the quantity of this debris and its component nutrients may help scientists
understand the health of the ecosystem and speed of forest carbon cycles and
more broadly, climate change.


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.The sampling guide divides the litter into 8 different functional groups
 (leaves, needles, twigs/branches, woody materials, seeds, flowers and other
 non-woody reproductive structures, other, and mixed material).The mass of each
 functional group is recorded with an accuracy of 0.01 grams.
 2.Sampling occurs at NEON sites that contain woody vegetation over 2 meters
 tall.
 3. Ground traps are sampled once per year.


## Obtain basic summaries of your data (Neonics)

5.  What are the dimensions of the dataset?


‘‘‘ r
#find the dimensions of the dataset
dim(Neonics)


## [1] 4623   30
```

```r
#returns number of rows and columns: 4623 rows by 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
#view a summary of the effect column to understand the most common effects that
#are studied
effects_summary <- summary(Neonics$Effect)
#sort from largest to smallest to display most common effects at the top
sorted_effects_summary <- sort(effects_summary, decreasing=TRUE)
print(sorted_effects_summary)
```

```
##       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
```

```
#three most common effects studied are population, mortality, and. behavior
#(in order)
```

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 were most studied (1803 and 1493 entries, respectively). I am guessing these are effects are of interest because neonicotinoids are intended to kill the insects, so these two effects can give direct insight into their effectiveness in that regard. Also, you cannot measure other behavior of the insects if they are not alive, so that may explain why the other behaviors are less frequently recorded.

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…]

```
#determine 6 most commonly studied species
#get a summary of the Species.Common.Name column
#use the maxsum argument, which Help says is an integer indicating how many
#levels should be shown for factors
species_summary <- summary(Neonics$Species.Common.Name, maxsum=6)
#sort from largest to smallest to display most common species at the top
sorted_species_summary <- sort(species_summary, decreasing=TRUE)
print(sorted_species_summary)
```

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

```
#Other is the most common result at 3196 observations, followed by honey bee
#667, parasitic wasp 285, tailed bumblebee 183, carnolian honey bee 152, bumble
#bee 140
```

Question: What do these species have in common? Why might they be of interest over other insects? > Answer: Almost all of these species are types of bumble bees. They might be of interest over other insects since they are pollinators.

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…]

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

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

```
#running the above command tells you it's a factor
```

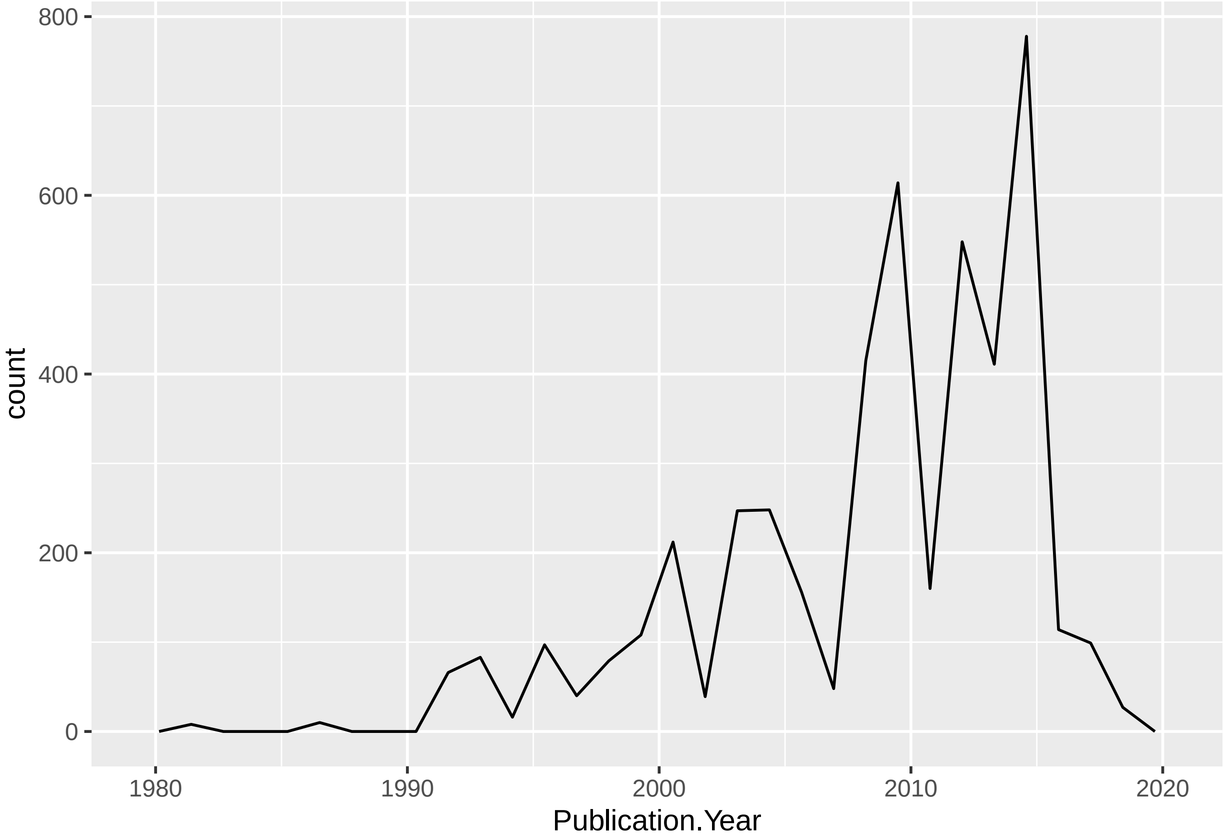
Answer:When looking at the data I can see there are some "/" in the Conc.1..Author column along with the numbers, therefore R processed it as factor data and not numeric.

Explore your data graphically (Neonics)

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

```
#generate a plot of studies per year with geom_freqpoly
studies_by_yr <- ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year))
plot(studies_by_yr)
```

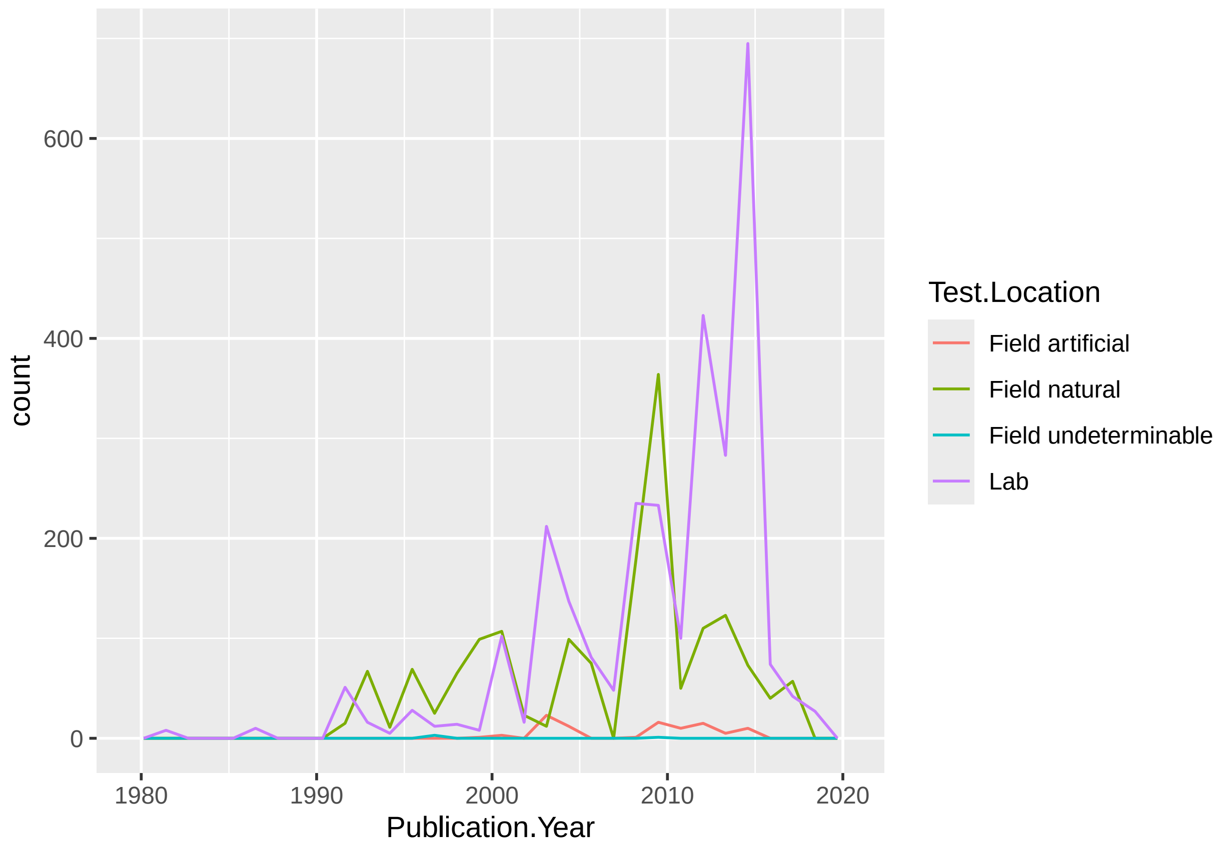
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
#reproduce the plot with a color aesthetic
studies_by_yr_colored <- ggplot(Neonics) + geom_freqpoly(aes(x =
Publication.Year, color = Test.Location))
#add color argument to geom_freqpoly function to add the color aesthetic
plot(studies_by_yr_colored)
```

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

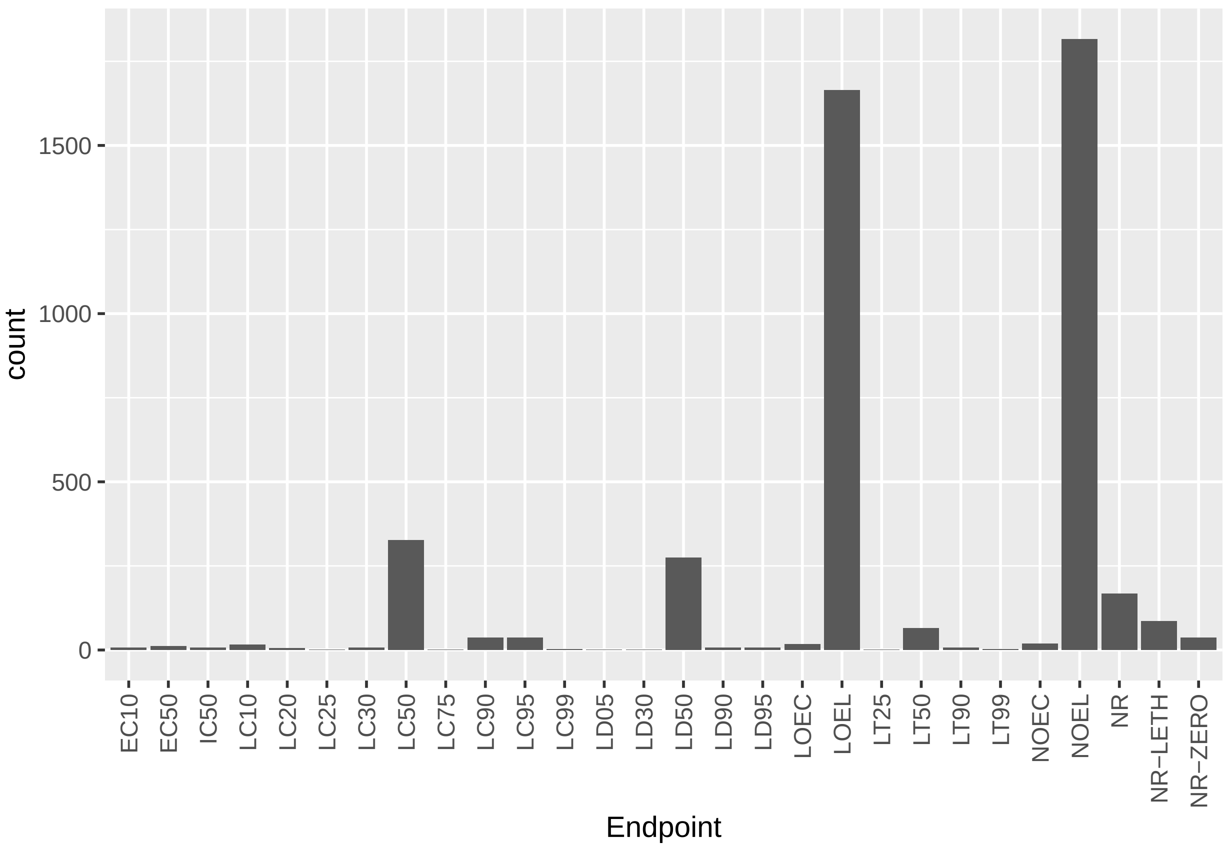


Interpret this graph. What are the most common test locations, and do they differ over time? > Answer:The most common test location is the lab for a short time in the early 1990s and then it is eclipsed by field natural until around 2000 when lab again becomes more popular until around 2008. From around 2008-2010 field natural is the most common test location. Then from 2010-2015 studies with lab test location skyrocket and it is far more common than any other location. From 2015-2020 lab studies dramatically decrease and are briedfly eclipsed by field natural studies. Overall, field natural and lab test locations are far more common than field artificial and field undererminable, and switch off which of the two of them is the most common over the 40 year period.

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…]

```
#create bar graph of Endpoint counts
endpoint_counts <- ggplot(data = Neonics, aes(x = Endpoint)) +
  geom_bar() + theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
  hjust=1))
endpoint_counts
```



```
#theme argument added to end of plot command to rotate and align X-axis labels
```

“ ” 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. LOEL means lowest-observable-effect-level, the lowest dose producing results that were significantly different from controls.NOEL means no-observable-effect-level, the highest dose producing effects not significantly different from control responses.

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.

```
#check the class of the collectDate column
class(Litter$collectDate)
```

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

```
#returns factor
#use Lubridate to work with dates since it's easier
library(lubridate)
Litter$collectDate <- ymd(Litter$collectDate) #put into Date class
class(Litter$collectDate) #confirms it is now in date format
```

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

```
#use unique to list the unique dates in the column
unique(Litter$collectDate)
```

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

```
#returns 2018-08-02 and 2018-08-30 so we know Litter was sampled on those dates
```

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

```
#create a new object to store unique plot IDs
unique(Litter$plotID)
```

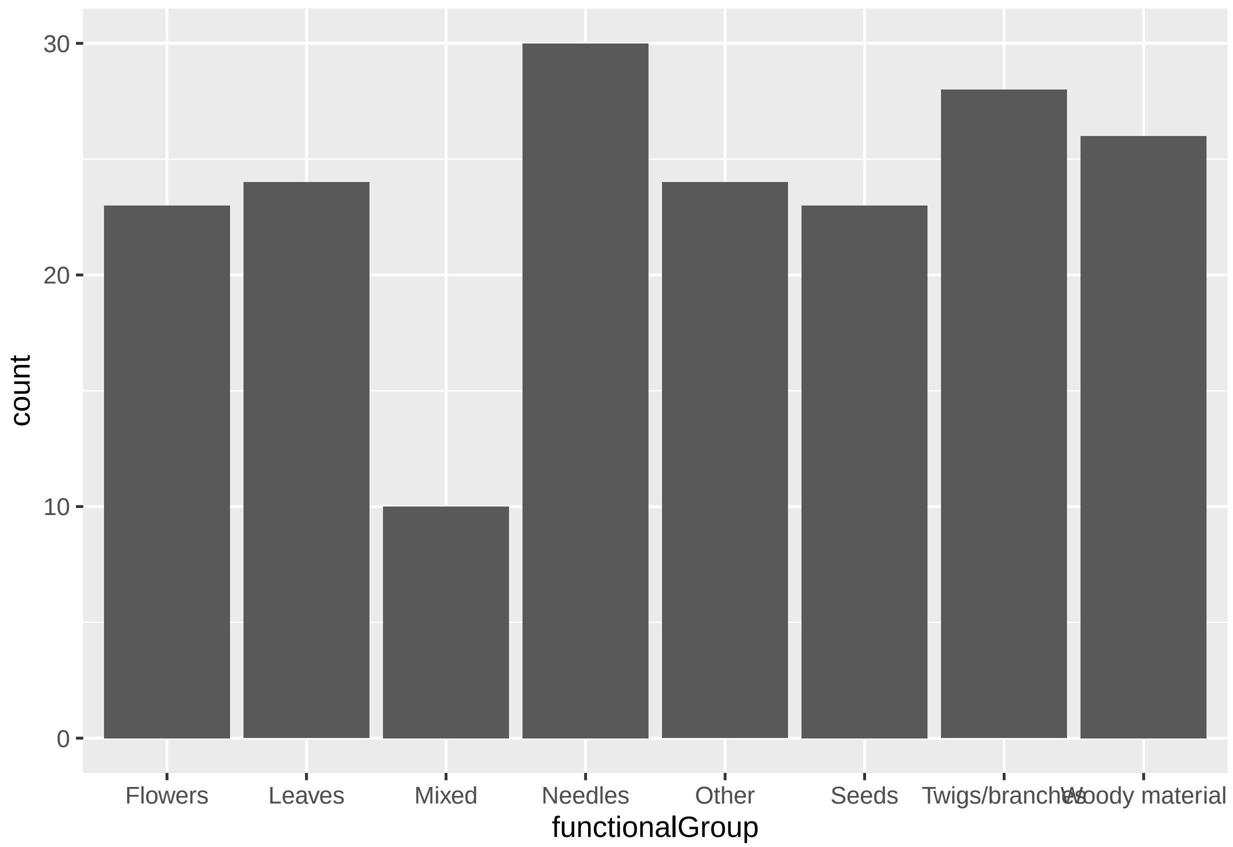
```
##  [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

```
#returns 12 different unique IDs: NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041
#NIWO_063 NIWO_047 NIWO_051 NIWO_058 NIWO_046 NIWO_062 NIWO_057
```

“ ” How is the information obtained from `unique` different from that obtained from `summary`? > Answer: The Summary function when applied to a column will give a count of the number of observations of each type. Whereas, the unique function will only return each unique observation, and will not count how many times it appears in the dataset.They both provide all unique observations but the Summary function goes beyond the unique function by also providing the count.

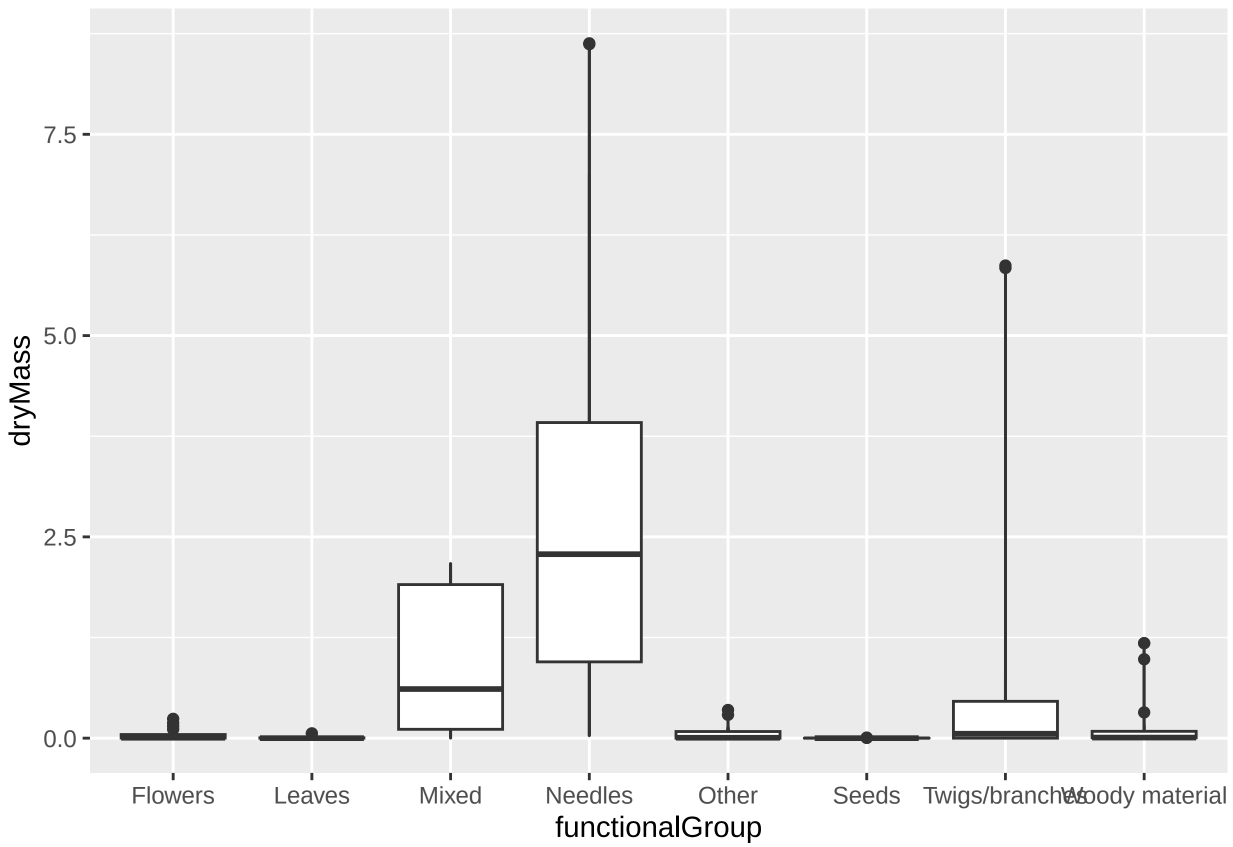
14. Create a bar graph of `functionalGroup` counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#create bar graph using ggplot and x is functional group column
ggplot(data = Litter, aes(x = functionalGroup)) + geom_bar()
```

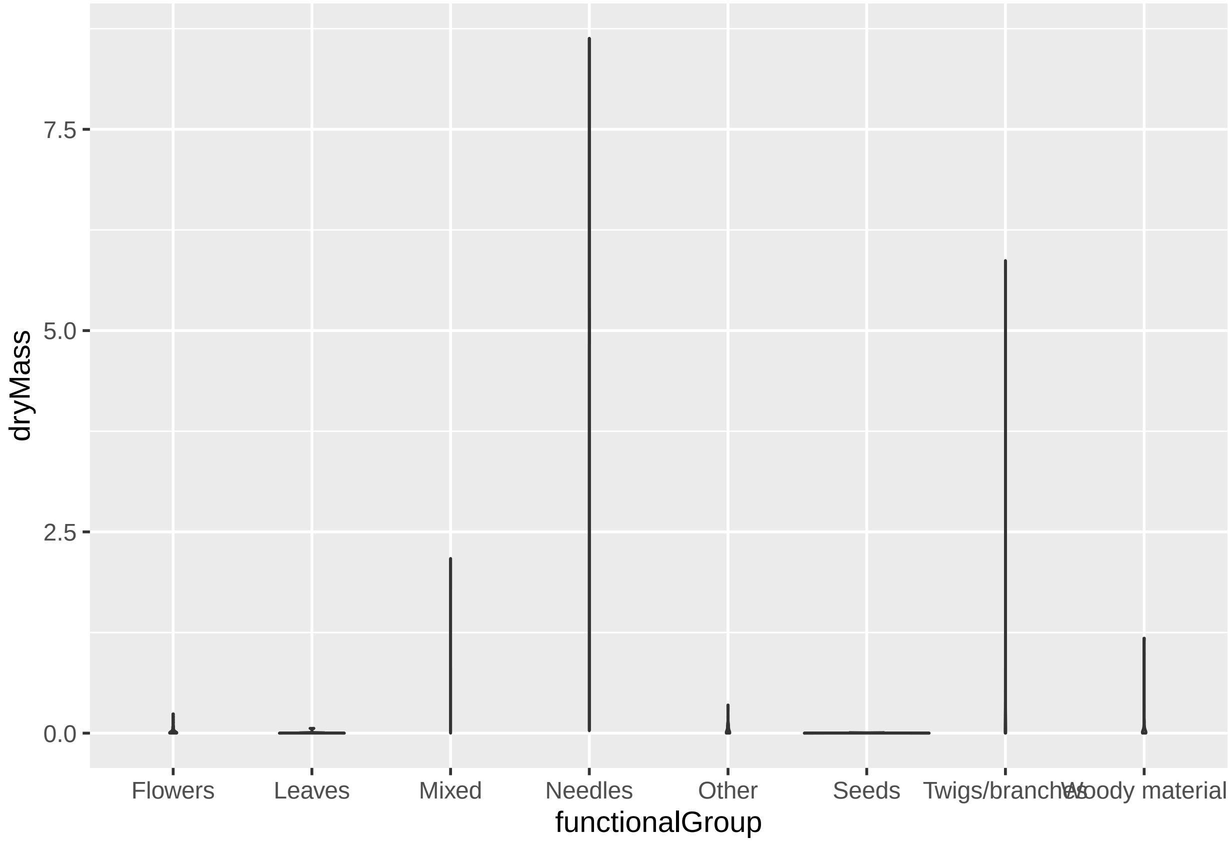


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of `dryMass` by `functionalGroup`.

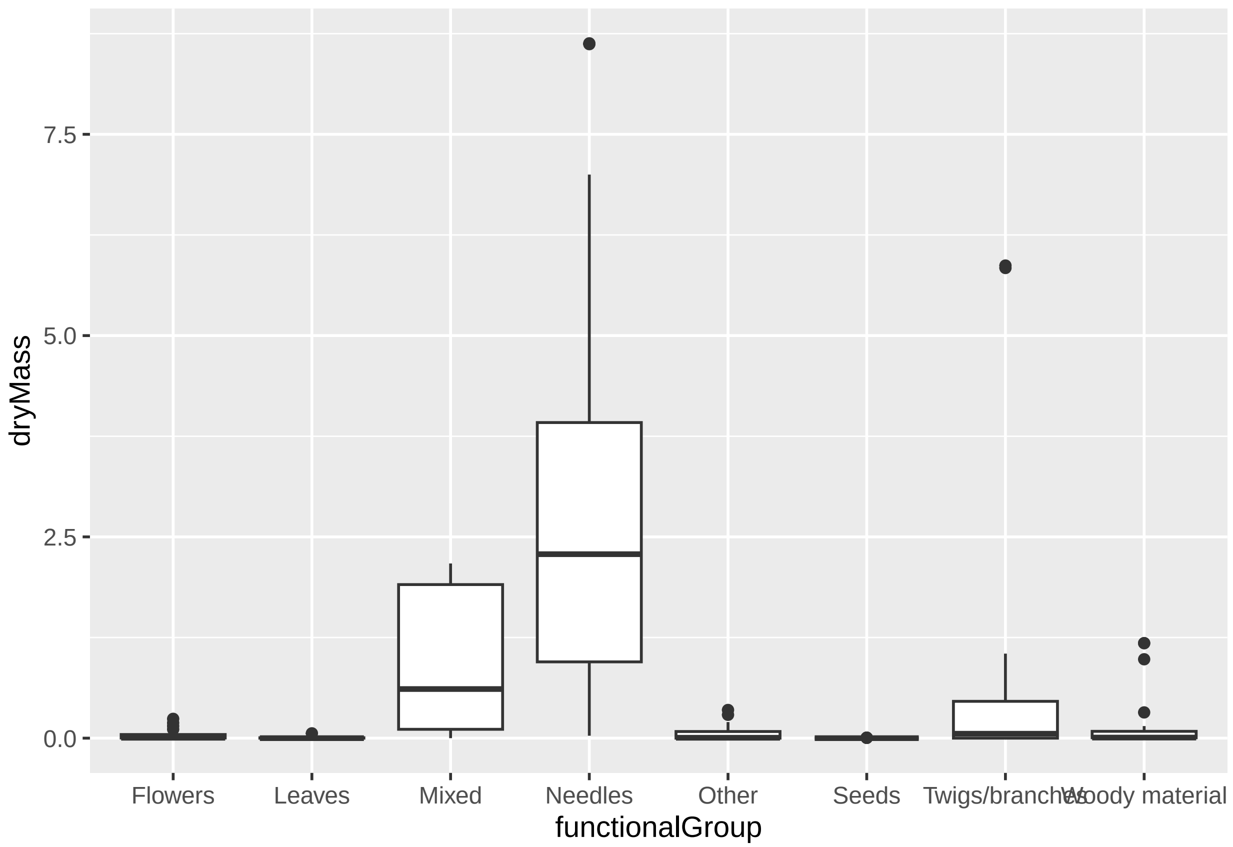
```
#again use ggplot with x=functional group column but add the two different plot
#types
#I am not clear if these are supposed to be on the same plot so here I am
#assuming they are
ggplot(data = Litter, aes(x = functionalGroup, y = dryMass)) + geom_violin() + geom_boxplot()
```



```
#violin plot only - this seems less useful
ggplot(data = Litter, aes(x = functionalGroup, y = dryMass)) + geom_violin()
```



```
#box plot only. I had to remove some NA values that were preventing the plot
#from displaying
ggplot(data = Litter, aes(x = functionalGroup, y = dryMass)) +
geom_boxplot(na.rm=TRUE)
```



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

> Answer:The boxplots are more effective here due to the disparity in number of observations per group. With some of the smaller groups, the shapes created by the violin plot become hard to read and distinguish from one another. The box plots show more in general, so they are more informative, aesthetically pleasing, and allow us to see the spread, median, IQR, and outliers more easily.

“ ” What type(s) of litter tend to have the highest biomass at these sites? > Answer:Needles have a much higher median biomass than any other type of litter. Mixed is the second highest litter type in terms of biomass, while all other litter types have a low biomass.