Assignment 3: Data Exploration

Gretchen Barbera

Fall 2024

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.

```
#install.packages("here")
library(here)
#install.packages("lubridate")
library(lubridate)
#install.packages("tidyverse")
library(tidyverse)
#install.packages("dplyr")
library(dplyr)
#install.packages("ggplot2")
library(ggplot2)
```

```
#install.packages("readr")
library(readr)
#step one is making sure I have all of the
#packages on my computer.
#I loaded the same packages from the lab 3 from last week

getwd()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```
here()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```
#Making sure that my working directory is
#putting stuff in the right place



#read.csv(
 # file = "/home/guest/EDE_Fall2024/Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv" ,
  #stringsAsFactors = TRUE)

#absolute path


Neonics <- read.csv(
  file = here("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"),
  stringsAsFactors = TRUE)

#View(Neonics)

Litter <- read.csv(
  file = here("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv"),
  stringsAsFactors = TRUE)

#View(Litter)
#relative path

#Figuring out the right path to upload my data was tricky but I
#eventually figured out how to get my path in.
#When I imported the data, R gave me the code library to follow
```

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: Researchers should be interested in the ecotoxicology of the neonicotinoids on insects because it can be used as a indicator for the overall health and the potential longterm threats of the ecosystem. Ecotoxins are a potential threat to the ecosystems in which they are used. Insects can be used as biodindicator species for what the overall health of the ecosystem looks like and help project what it might look like in the future. The environmental and health impacts of the neonicotinoids on insetcs could be used to demonstrate the possibilities for bioaccumulation through the ecosystem and could help project potential concerns in larger parts of the ecosystem. It is also important to see an environmental disturbance that could potential occur from the ecotoxins- the potential threat to impact not just the insect pest targets, but also other insects within the ecosystem is strong. For example, beneficial bugs such as pollinators could be affected by the neonicotinoids would could leave a detrimental impact to an ecosystem. Studying and understanding the nrelationship between the ecotoxicology of the neonicotinoids and it's impact on insects could be used to project future environmental and ecosystem detriments.

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: Studying litter and woody debris in forests, particularly in the context of the Niwot Ridge long-term ecological research (LTER) station, is valuable in understanding an ecosystem and predicting long-term trends. Litter and woody debris can serve as a habitat for insects, small mammals, fungi, and more. They can also be an indicator of the overall health of the ecosystem as they are part of the nutrient cycling process. As they decompose, they release nutrients back into the soil that are essential in supporting plant growth ecosystem health. Litter and woody debris are also indicators of ecosystem resilience and can provide insights into how forests respond to disturbances, such as storms or insect outbreaks, and their capacity for recovery. Collecting the data would help researchers predict trends in litter and debris accumulation, decomposition rates, and their responses to environmental changes, such as 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. Litter and woody debris are collected from elevated and ground traps. In areas where the vegetation is primarily forests, litter samples focus on 20 plots measuring 40m X 40m 2. In low-saturated vegetation areas, or areas that have smaller/lower to the ground vegetation, the sampling consists of 4 plots, with 40m X 40m measurments.
   3. "Trap placement within the may be targeted, or randomized depending on the vegetatioation". In sites where there is less than 50% aerial vegetation cover the litter traps are randomly placed within the plot.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

```r
#dim is used to find the dimensions of the dataset
#4623 rows and 30 columns

# dim(Litter)
```

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...]

```r
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
#summarizing the effect column on the dataset so
#I can get a better understanding of the dataset
effect_counts <- table(Neonics$Effect)
#i created a table for the summarized
#data that i found above. Since it gave me a
#character and not just a number
#I needed to see how many times the data mentions
#one of the effects
sorted_effects <- sort(effect_counts, decreasing = TRUE)
#using the sorted effects that i got above
#I just sorted the effects from most amount of values to least amount of values
#(most common to least common in the dataset)
sorted_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
#variable name that I created. Using this, I can now view a table of the effect
#and the corresponding amount of times it is mentioned in the data - magnitude
```

Answer: The most common effects that are studied are population, mortality, and behavior. These would be of particular interest because the population size and the mortality rate, are great indicators of ecosystem health. It is also interesting to note the behavior of the insects in conjunction with the population size and mortality rate- this could be used to indicate where there is an increase in mortality within 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...]

```
Neonics$Species.Common.Name <- as.factor(Neonics$Species.Common.Name)
#I am looking for the different types of species mentioned in the data
#by looking for the common name

Species.Common.Name_summary <- summary(Neonics$Species.Common.Name)
# I am getting the summary of the Species Common Name in the table

Species.Common.Name_summary_df <- as.data.frame(Species.Common.Name_summary)

top_six_species <- head(sort(Species.Common.Name_summary, decreasing = TRUE), 6)
#I named it the top_six_species
#I am getting the species counts in decreasing order so the most commonly
#studied species will be first and then so on
print(top_six_species)
```

```
##               (Other)             Honey Bee        Parasitic Wasp
##                   670                   667                   285
## Buff Tailed Bumblebee   Carniolan Honey Bee            Bumble Bee
##                   183                   152                   140
```

Answer: Honeybee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and "other". They are all pollinators. This would be of interest in the study because as they go from plant to plant, they may interact with the insecticides. Many neonicotinoids are known to affect pollinators and since pollinators are essential in plant reproduction, understanding the impacts the neonicotinoids could have on pollinators could be an indicator for ecosystem health.

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...]

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

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

```
#class= factor.
```

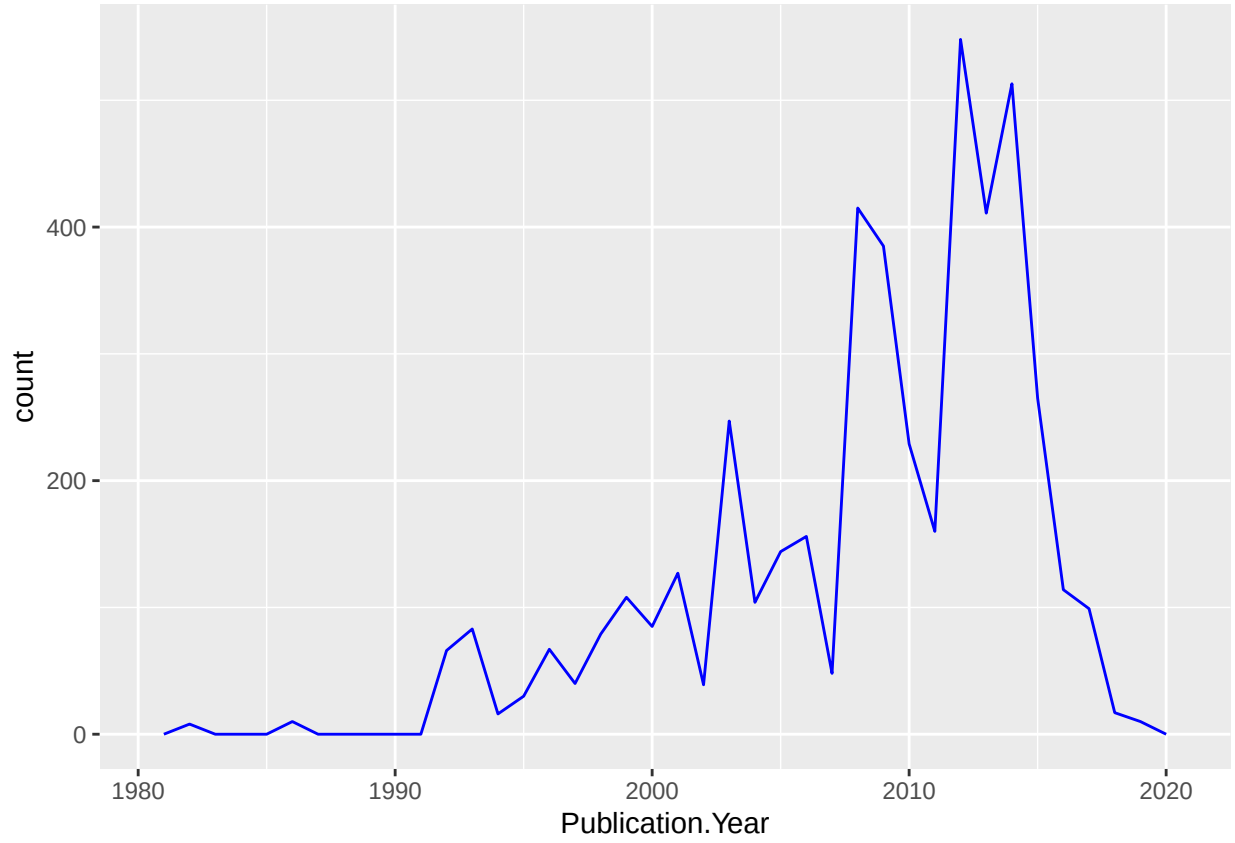
Answer: Factor. That means that there is a mix of data types in the column, 'Con.1..Author'. When looking at the data manually, I noticed that there were numeric values and also NR which could mean that the data was not recorded or not relevant. Either way, it changes the data so it is categorical- so the NR value is now its own 'unique' data.

Explore your data graphically (Neonics)

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

```
Neonics$Publication.Year <- as.numeric(as.character(Neonics$Publication.Year))
#when I had to convert the date into numeric data so it could be plotted

ggplot(Neonics, aes(x = Publication.Year)) +
  geom_freqpoly(binwidth = 1, color= "blue")
```

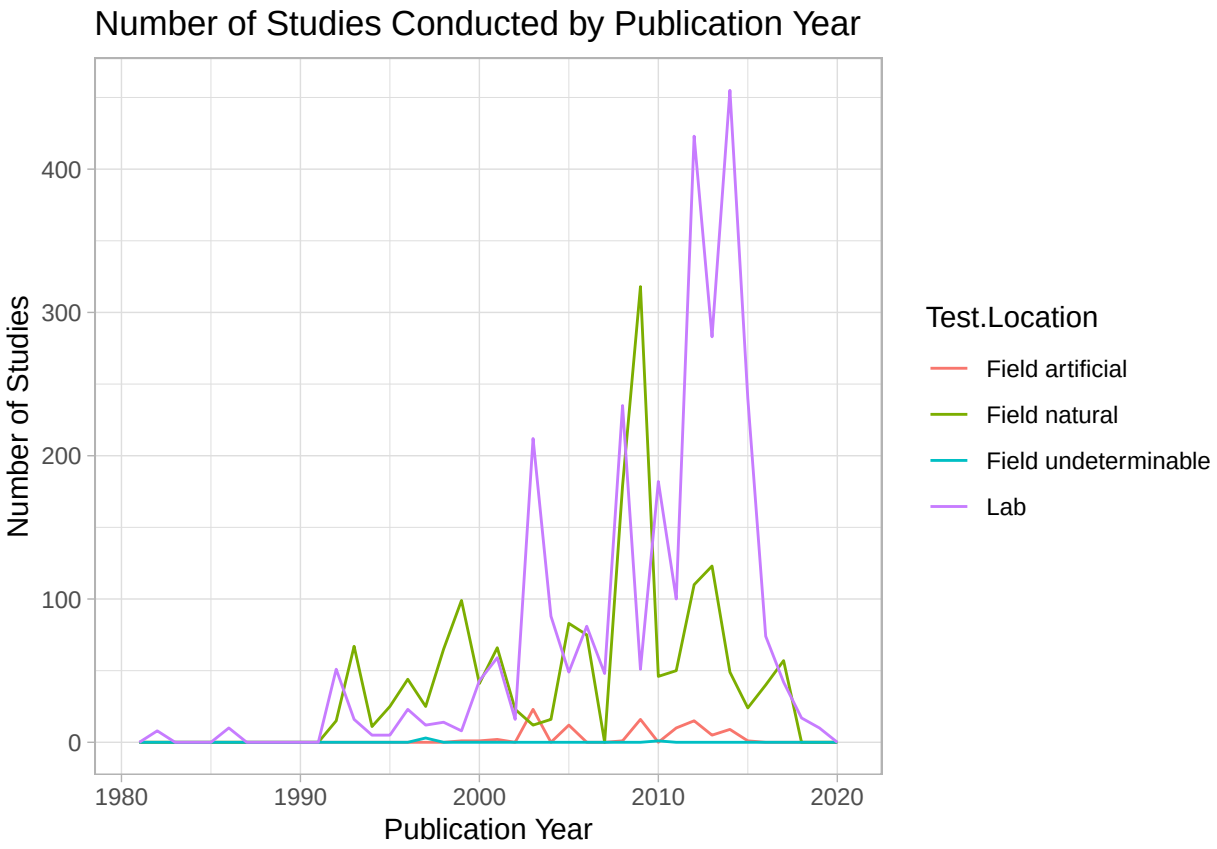



```
#binwidth is the frame of the graph
#color is blue
```

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

```
Neonics$Test.Location <- as.factor(Neonics$Test.Location)
#make sure that the Test.Location is a factor

ggplot(Neonics, aes(x = Publication.Year, colour = Test.Location)) +
  geom_freqpoly(binwidth = 1) +
  labs(title= "Number of Studies Conducted by Publication Year",
       x= "Publication Year",
       y= "Number of Studies") +
  theme_light()
```




```
# labs adds a title to make it look nice and so I can label the plot
#I had the test locations show up as colors but putting it within
#the aesthetic ()
#light theme so the colors look better
```

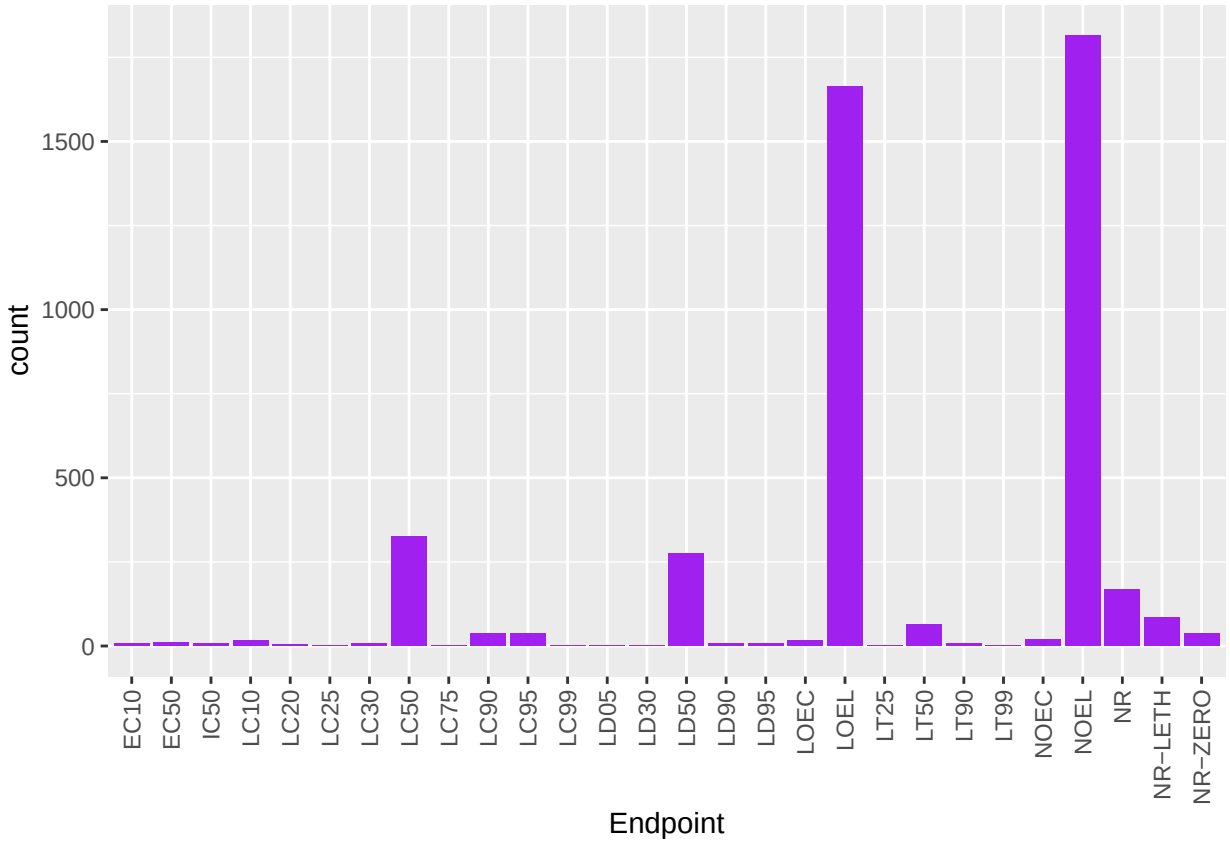
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common testing locations are the lab and the field Natural sites, while there are a few field artificial sites, they are no wear near the numbers as the other twon. Field undeterminable remains a straight line along the x axis displaying 0s in the number of studies. Towards the end of the 2000s heading into the 2010s, the 'field Natural' was the most common testing location. Between 2010 and 2020 the 'lab' was the most common testing location showing a spike in the number of studies around 2015.

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, aes(x= Endpoint)) +
  geom_bar(fill = "purple") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: NOEL and LOEL. LOEL stands for the lowest-observed- effect levels which means the lowest dose or concentration producing effects that were significantly different from the response controls (Ecotox_codeAppendix) NOEL stands for No-Observable-Effect-Level which is the highest dose or concentration that produces effects that are not significantly different from the controls (Ecotox_CodeAppendix)

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)
```

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

```
#Entered with the class 'factor'

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

class(Litter$collectDate)
```

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

```r
#changed the class so now it is 'Date'


august_2018_samples <- Litter$collectDate[Litter$collectDate
                                          >=  as.Date("2018-08-01") &
                                            Litter$collectDate
                                          <= as.Date("2018-08-31")]

unique_august_2018_samples <- unique(august_2018_samples)

print(unique_august_2018_samples)
```

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

```r
#there are only 2 unique dates and they
#were collected on August 2nd and August 30th.
```

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
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
```

```r
#gives me how many entries there are in the plotID

length(unique(Litter$plotID))
```

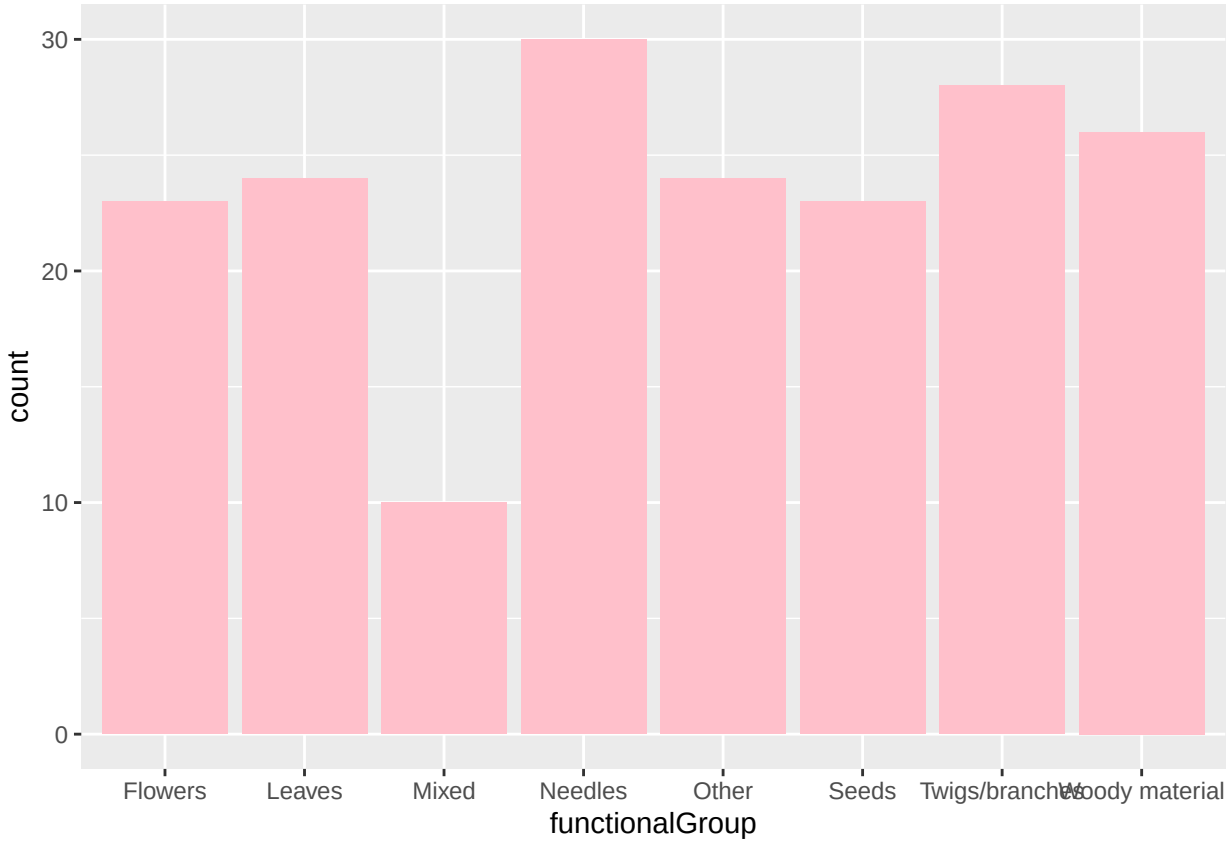
```
## [1] 12
```

```r
#how many unique sites that are sampled
```

Answer: The 'unique' function helps me understand what the unique values are in a column- it doesn't give me a further chraracteristic of the data, while the "summary" function told me the species and how many times they were studied. The 'unique' function helped me understand the distribution and characteristics of the data (how many plots were studied) while the 'summary' function gave me the definitive number of a specific characteristic that I was looking for- in this case the top six species and their corresponding values of how many times they were studied. The 'summary' function in this data would tell me the number of plots and how many times there were sampled.

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(fill = "pink")
```

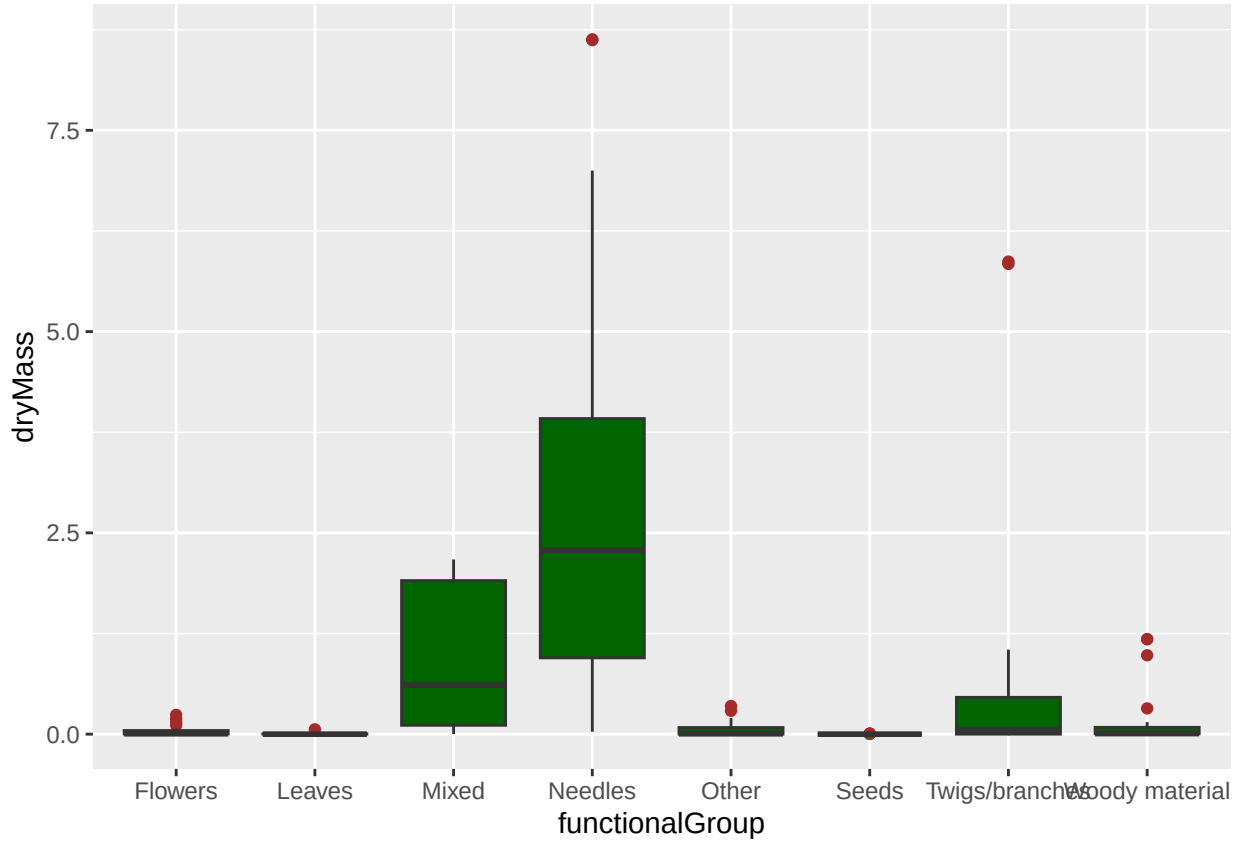



```
#Litter types in pink. We see that  'Needles' and 'Twigs and branches'
#functionalGroups are counted the most
```

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

```
boxplot <- ggplot(Litter, aes(x= functionalGroup, y= dryMass)) +
  geom_boxplot(fill = "darkgreen", outlier.color = "brown")

print(boxplot)
```



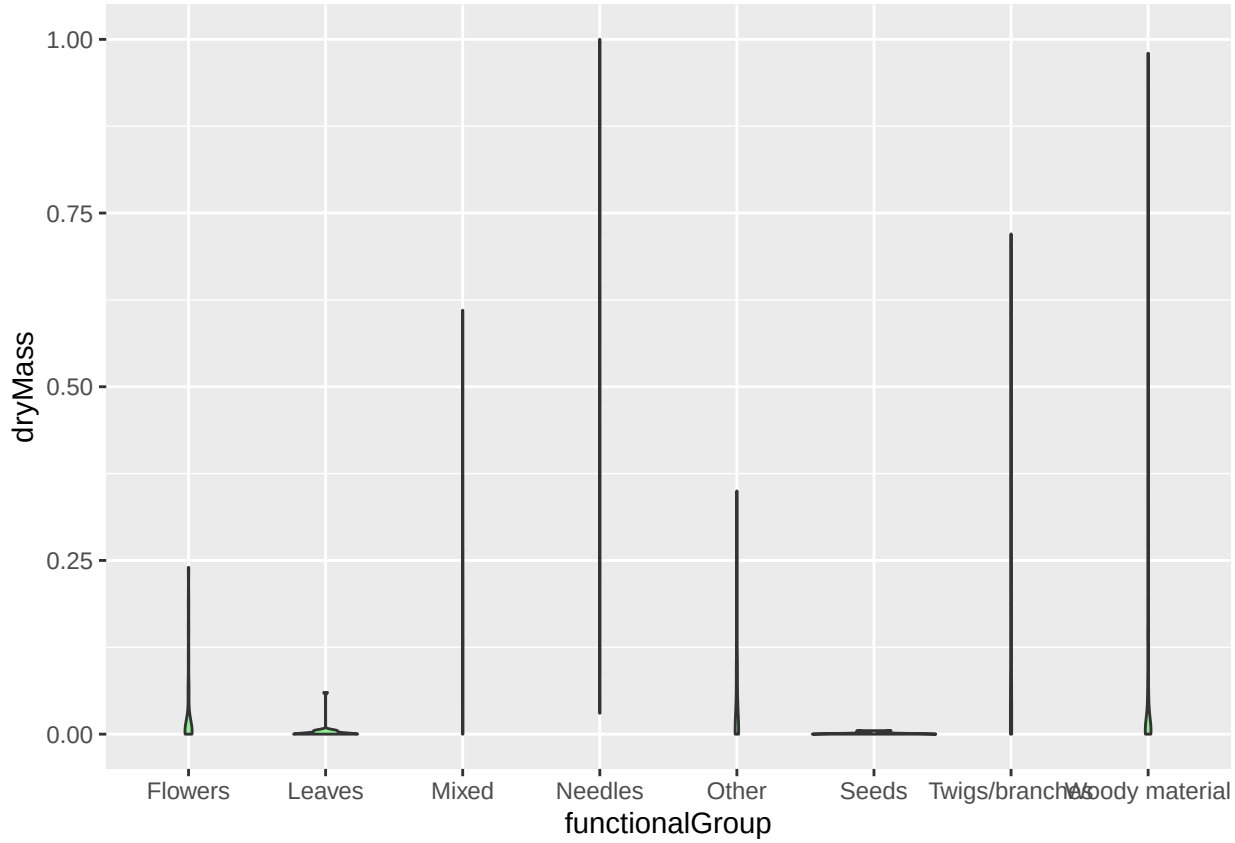

```
#showing the outliers in a different color so I can really see the difference
#in the data


#it comes out as just lines which could mean that most of the data is in
#the 75 percentile or that the dryMass values are smaller than the frame of
#the violin, which tells me I should find a new
#range and set a limit to the dryMass

violin_plot <- ggplot(Litter, aes(x= functionalGroup, y = dryMass)) +
  geom_violin(fill = "lightgreen") +
  ylim(0,1)

print(violin_plot)
```

```
## Warning: Removed 30 rows containing non-finite outside the scale range
## (‘stat_ydensity()‘).
```




```
#median(Litter$dryMass)
#I have to find the median so I know how to set my range

#violin_plot <- ggplot(Litter, aes(x= functionalGroup, y = dryMass)) +
  #geom_violin(fill = "lightgreen") +
  #ylim(0,0.015)

#print(violin_plot)

#this shows me a better visualization- however, this data range does not account
#for the dryMass values that are higher than my limit (0.015) so it doesn't show
#me the whole data set/ the types of litter with the highest biomass

#ggplot(Litter, aes(x= dryMass)) +
  #geom_histogram(binwidth = .1)

  #I plotted a histogram because I was confused about the lines and
#it shows that most of the dryMass count comes out as 0 or around 0.
```

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

> Answer: The boxplot was more effective in visualizing the data because gave me a clear median as well as the range of the dryMass from the functionalGroup. It was also able to show me tific he outliers within the data clearly so I can see if there is any deviation from the median/if the points are significantly different from the rest of the data collected. I needed to change the range

of the dryMass on my violin plot so it would zoom in on my data- without adjusting the range, it just showed me lines which weren't helpful in helping me understand the data/wasn't a good visualization of what the data looked like. The boxplot helped me visualize and target which functionalGroup had the largest dryMass from the start. pec What type(s) of litter tend to have the highest biomass at these sites?

Answer: 'Needles', 'twigs/ branches', and 'woody materiab sl' la have the highest biomass at these sites. This makes sense after reading the NEON_Litterfall_UserGuide for question 3- where they detail that litter and woody debris are what they are mostly sampling for.