Assignment 3: Data Exploration

Ina Liao

Fall 2023

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 Sakai.

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. Check your working directory, load necessary packages (tidyverse, lubridate), 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
#install.packages("tidyverse")
library(tidyverse)
#install.packages("lubridate")
library(lubridate)
#install.packages("dplyr")
library(dplyr)
```

```r
getwd()
```

```
## [1] "/Users/inaliao/Desktop/EDE_Fall2023"
```

```r
setwd("/Users/inaliao/Desktop/EDE_Fall2023/Assignments")
getwd()
```

```
## [1] "/Users/inaliao/Desktop/EDE_Fall2023/Assignments"
```

```
Neonics<-read.csv("/Users/inaliao/Desktop/EDE_Fall2023/Assignments/ECOTOX_Neonicotinoids_Insects_raw.cs
Litter<-read.csv("/Users/inaliao/Desktop/EDE_Fall2023/Assignments/NEON_NIWO_Litter_massdata_2018-08_raw
```

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 used to protect crops from insect pests.; however, it might have intended impacts on non-target insects and other organisms. Researching the ecotoxicology of neonicotinoids could help us understand the full picture of neonicotinoids' impacts on 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.

   Answer: When litter and woody decompose, they release nitrogen, carbon, and other critical nutrients into the soil, improving soil health and enriching forest ecosystems.

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: Litter and woody debris data were collected through spacial sampling: the sampling was executed at terrestrial NEON sites that contain woody vegetation greater than 2 meters tall. Spatial random sampling has the following advantages: 1. reducing the measurement error in the data 2. reducing the needed data size for robust analysis 3. information tends to be more generalizable

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
View(Neonics)
num_dimention<-ncol(Neonics);num_dimention
```

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

   Answer:There are 30 dimentions of the dataset.

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?

```
unique(Neonics$Effect) #to get the unique values in the "Effect" column
```

```
##  [1] "Mortality"        "Growth"           "Population"       "Immunological"
##  [5] "Cell(s)"          "Behavior"         "Reproduction"     "Development"
##  [9] "Genetics"         "Enzyme(s)"        "Feeding behavior" "Avoidance"
## [13] "Intoxication"     "Biochemistry"     "Hormone(s)"       "Accumulation"
## [17] "Morphology"       "Histology"        "Physiology"
```

```
Neonics$Effect<-as.factor(Neonics$Effect) #convert data type: character to factor
count_effect<-summary(Neonics$Effect) #calculate frequency
```

```
count_effect_top<-sort(count_effect,decreasing=TRUE) #sort the frequency descendingly
head(count_effect_top, n=6)
```

```
##       Population        Mortality         Behavior Feeding behavior 
##             1803             1493              360              255 
##     Reproduction      Development 
##              197              136 
```

Answer:The most common effect that is studied is population (n=1803), followed by mortality (n=1493), behavior (n=360), feeding behavior (n=255), reproduction (n=197), and development (n=136).

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: The `sort()` command can sort the output of the summary command…]

```
Neonics$Species.Scientific.Name<-as.factor(Neonics$Species.Scientific.Name) #convert data type: charact
count_species<-summary(Neonics$Species.Scientific.Name) #calculate frequency
count_species_top<-sort(count_species,decreasing=TRUE) #sort the frequency descendingly
head(count_species_top,n=7)
```

```
##                       (Other)                Apis mellifera 
##                           974                           667 
##             Bombus terrestris   Apis mellifera ssp. carnica 
##                           183                           152 
##              Bombus impatiens Apis mellifera ssp. ligustica 
##                           140                           113 
##             Popillia japonica 
##                            94 
```

Answer:The six most studied species are Apis mellifera (n=667), Bombus terrestris (n=183), Apis mellifera ssp. carnica (n=152), Bombus impatiens (n=140), Apis mellifera ssp. ligustica (n=113), and Popillia japonica (n=94).

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?

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

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

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

```
## [1] "27.2" "19.7" "47"   "25"   "13"   "268" 
```

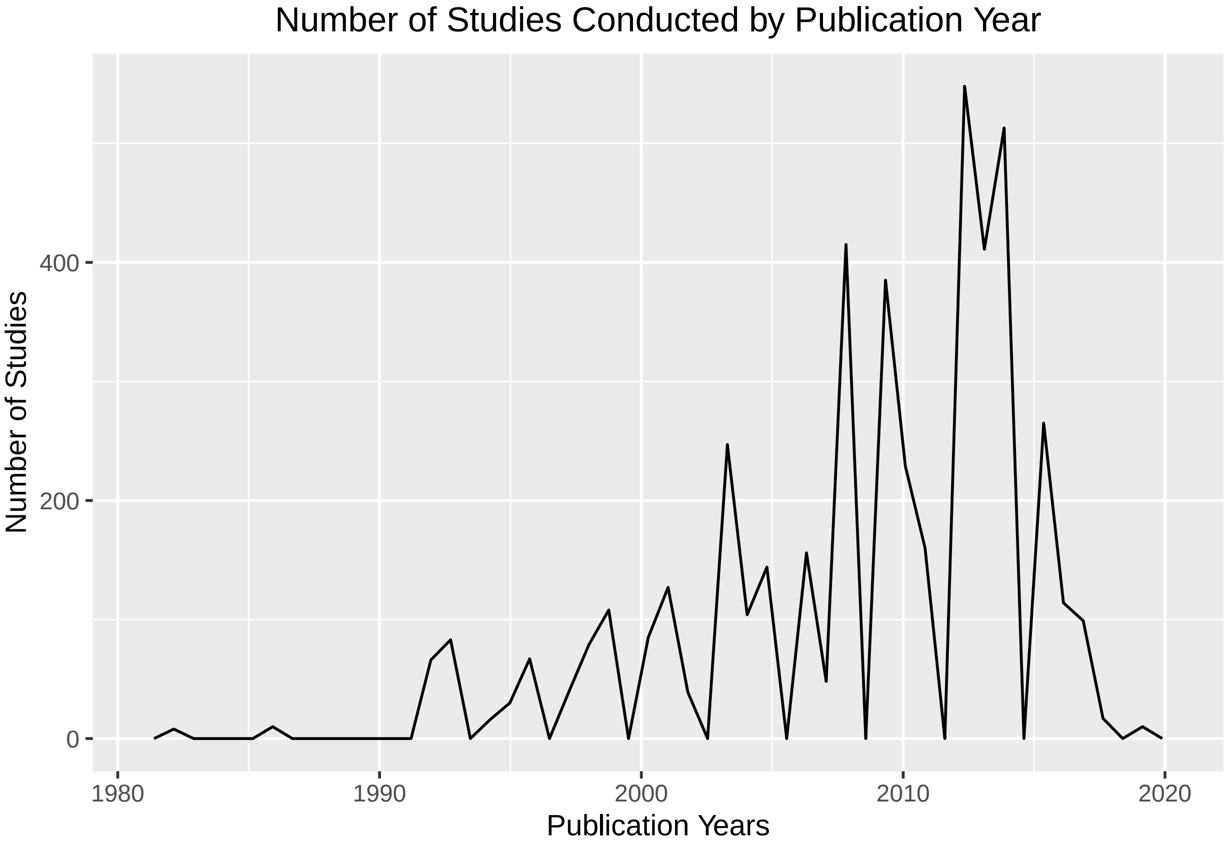
Answer:Because the values contained “ ” in the column.

Explore your data graphically (Neonics)

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

```
freq<-ggplot(data=Neonics)+
  geom_freqpoly(aes(x=Neonics$Publication.Year),bins=50,show.legend =TRUE)+
  theme(legend.position="top",plot.title = element_text(hjust = 0.5))+
  labs(title="Number of Studies Conducted by Publication Year",x="Publication Years",y="Number of Studi
freq
```

```
## Warning: Use of `Neonics$Publication.Year` is discouraged.
## i Use `Publication.Year` instead.
```



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

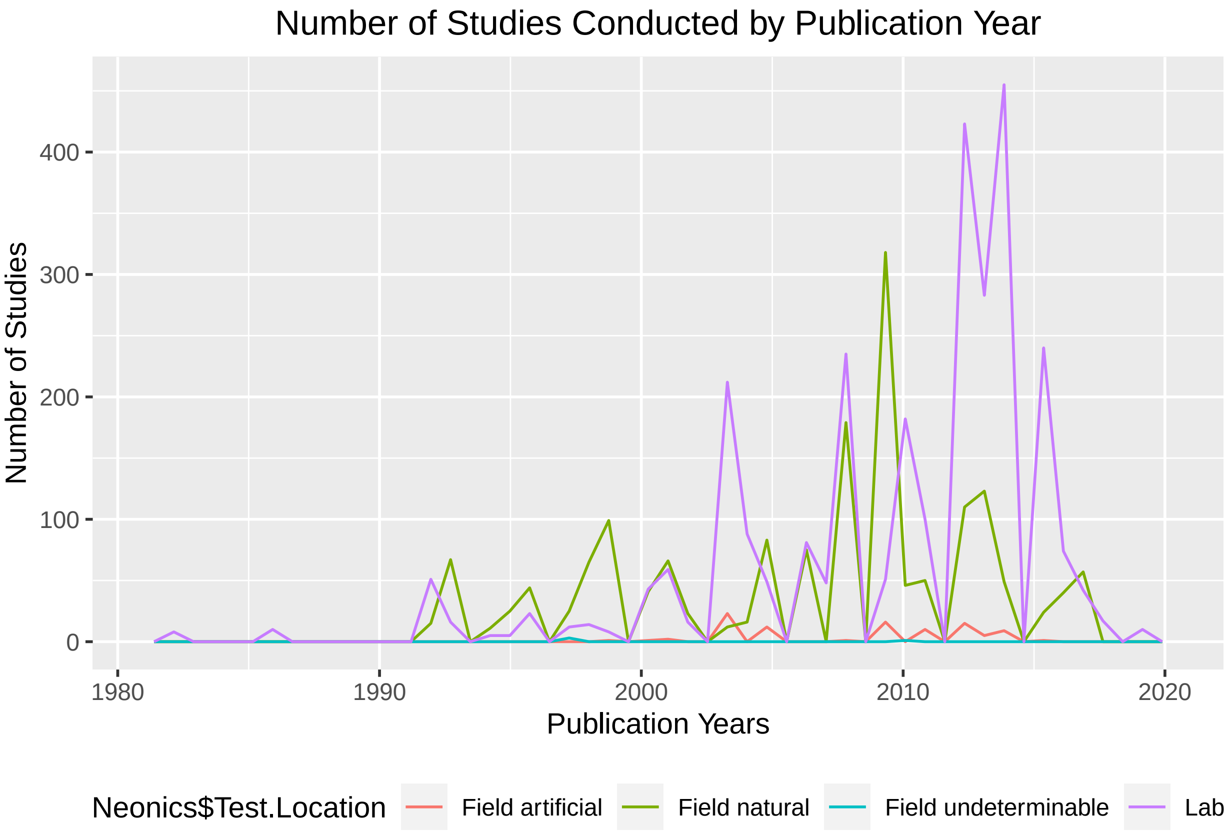
```
unique(Neonics$Test.Location)
```

```
## [1] "Lab"                   "Field natural"        "Field artificial"
## [4] "Field undeterminable"
```

```
freq_location<-ggplot(data=Neonics)+
  geom_freqpoly(aes(x=Neonics$Publication.Year,color=Neonics$Test.Location),bins=50,show.legend=TRUE)+
  theme(legend.position="bottom",plot.title = element_text(hjust = 0.5))+
  labs(title="Number of Studies Conducted by Publication Year",x="Publication Years",y="Number of Studi
freq_location
```

```
## Warning: Use of `Neonics$Publication.Year` is discouraged.
## i Use `Publication.Year` instead.

## Warning: Use of `Neonics$Test.Location` is discouraged.
## i Use `Test.Location` instead.
```



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

> Answer:Before 2010, the natural field was the most common test location. However, the lab replaced it, being the most common test location since 2010. The number of studies conducted in the lab and nature field had decreased to 0 by the end of 2020.

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.

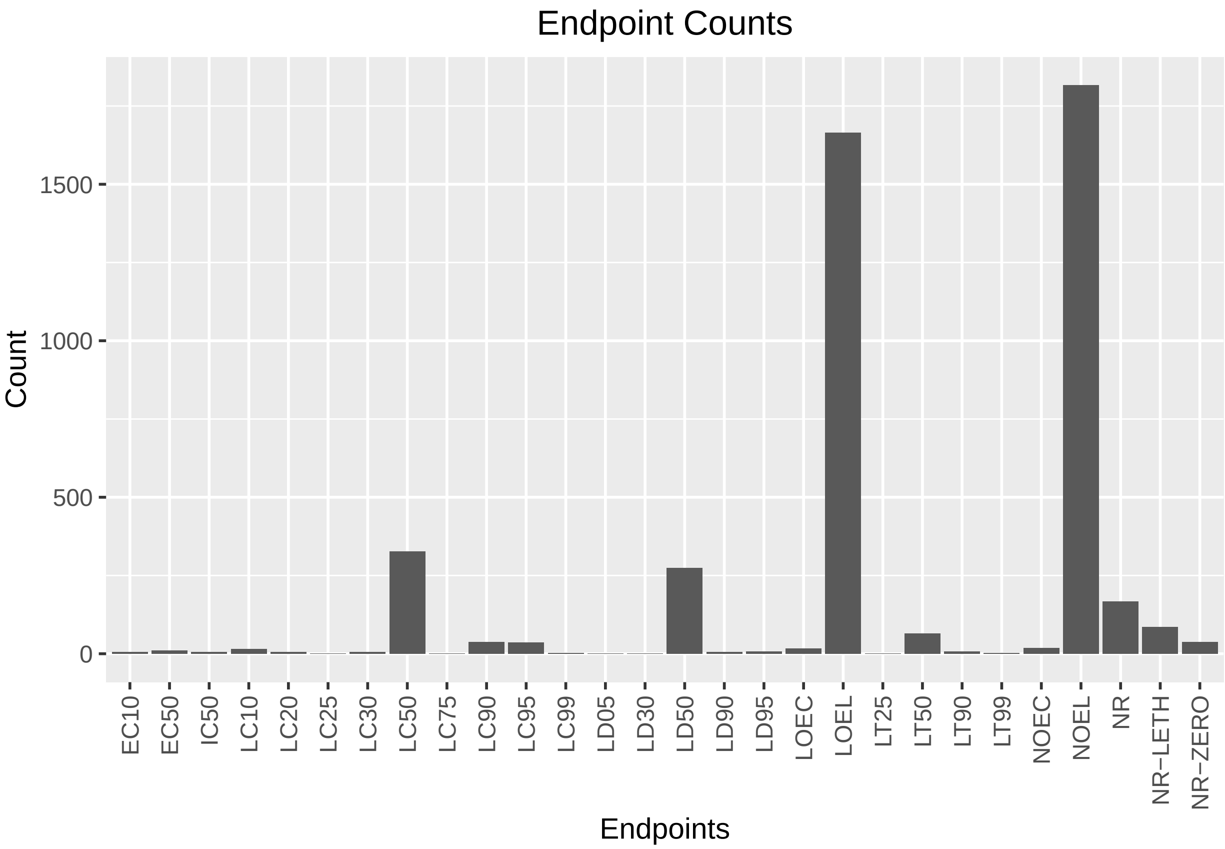
[**TIP**: Add `theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))` to the end of your plot command to rotate and align the X-axis labels…]

```
bar_endpoint<-ggplot(data=Neonics)+
  geom_bar(aes(x=Neonics$Endpoint),bins=50,show.legend=TRUE)+
  theme(legend.position="top",axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1),plot.title =
  labs(title="Endpoint Counts",x="Endpoints",y="Count")
```

```
## Warning in geom_bar(aes(x = Neonics$Endpoint), bins = 50, show.legend = TRUE):
## Ignoring unknown parameters: `bins`
```

```
bar_endpoint
```

```
## Warning: Use of `Neonics$Endpoint` is discouraged.
## i Use `Endpoint` instead.
```



Answer: NOEL and LOEC are the two most common endpoints. NOEL is defined as no-observable-effect-level, which is the terrestrial that has the highest producing effects and is not significantly different from responses of controls according to the author's reported statistical test. LOEC is the aquatic with the lowest observable effect concentration.

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.

```
View(Litter)
class(Litter$collectDate) #determine the class of collectData
```

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

```
collect_date<-ymd(Litter$collectDate) #change the data type to date
class(collect_date)
```

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

```
Aug_dates<-unique(collect_date)
Aug_dates #August 2th and 30th were the sample litter in 2018
```

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

13. Using the `unique` function, determine how many plots were sampled at Niwot Ridge. How is the information obtained from `unique` different from that obtained from `summary`?

```
num_plot<-unique(Litter$plotID)
num_plot
```

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

Answer:There were 12 plots sampled at Niwot Ridge. By using the unique function, we can only know the non-duplicated values while using the summary function, we can retrieve the frequency of each unique plot.

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.

```
unique(Litter$functionalGroup)
```

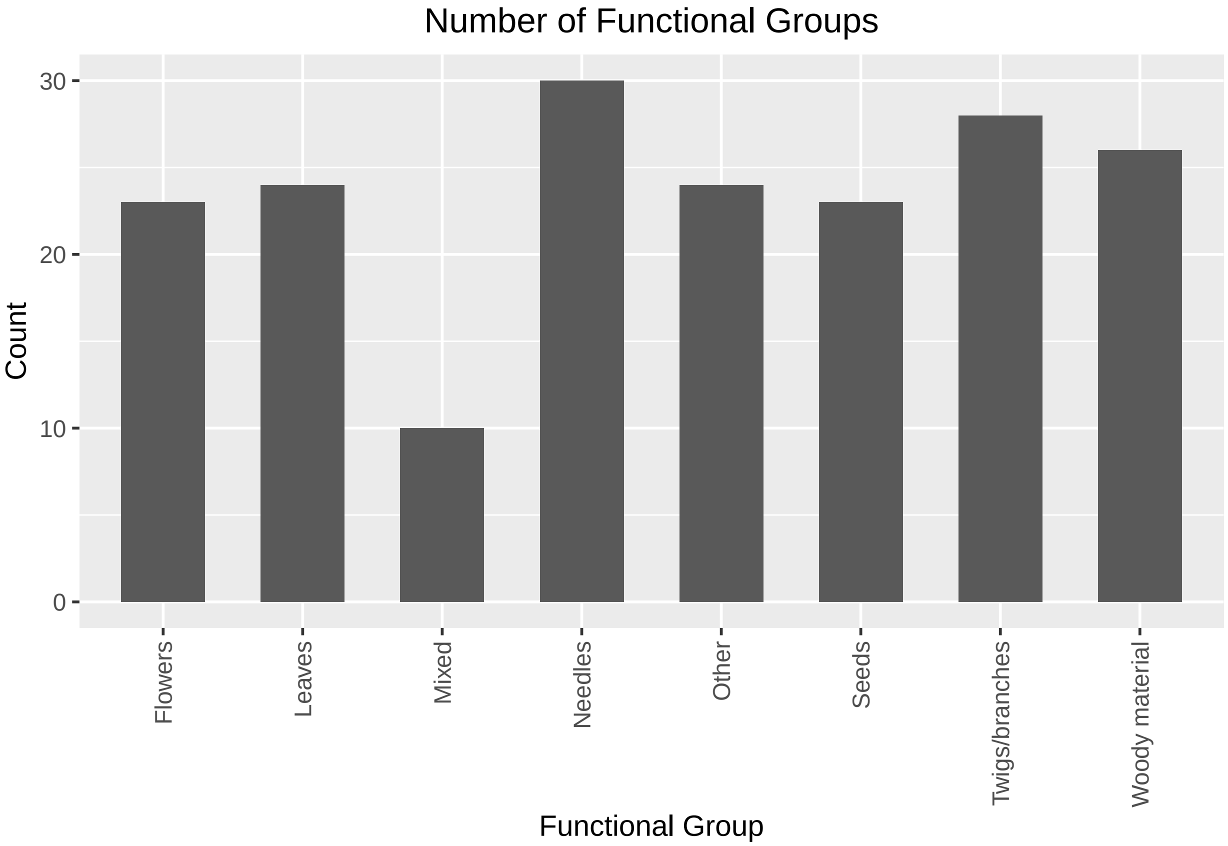
```
## [1] "Twigs/branches" "Seeds"          "Woody material" "Flowers"
## [5] "Needles"        "Other"          "Leaves"         "Mixed"
```

```
bar_func<-ggplot(data=Litter)+
  geom_bar(aes(x=Litter$functionalGroup),bins=50,show.legend=TRUE,width=0.6)+
  theme(legend.position="top",axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1),plot.title =
  labs(title="Number of Functional Groups",x="Functional Group",y="Count")
```

```
## Warning in geom_bar(aes(x = Litter$functionalGroup), bins = 50, show.legend =
## TRUE, : Ignoring unknown parameters: `bins`
```

```
bar_func
```

```
## Warning: Use of `Litter$functionalGroup` is discouraged.
## i Use `functionalGroup` instead.
```

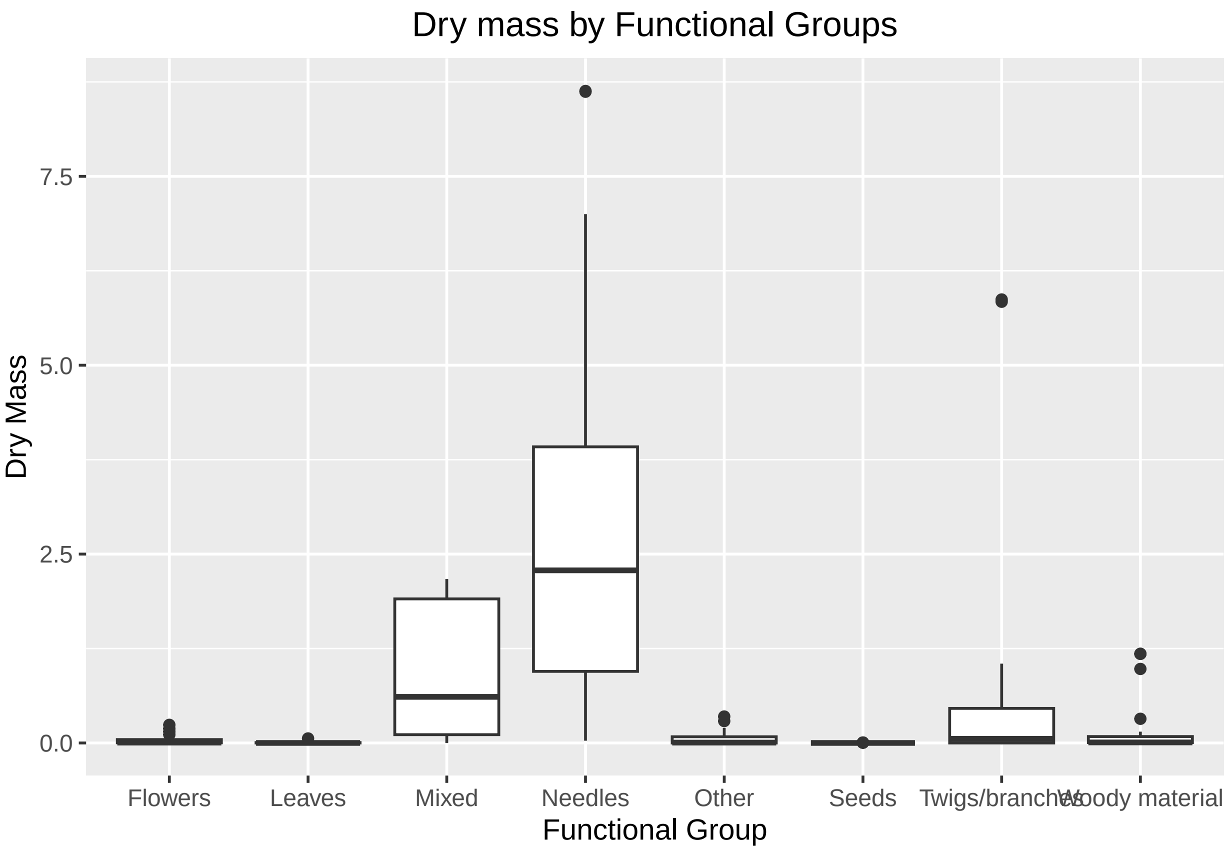


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```r
box_drymass<-ggplot(data=Litter)+
  geom_boxplot(aes(x=Litter$functionalGroup,y=Litter$dryMass),show.legend=TRUE)+
  theme(legend.position="bottom", plot.title=element_text(hjust=0.5))+labs(title="Dry mass by Functional
box_drymass
```

```
## Warning: Use of `Litter$functionalGroup` is discouraged.
## i Use `functionalGroup` instead.
```

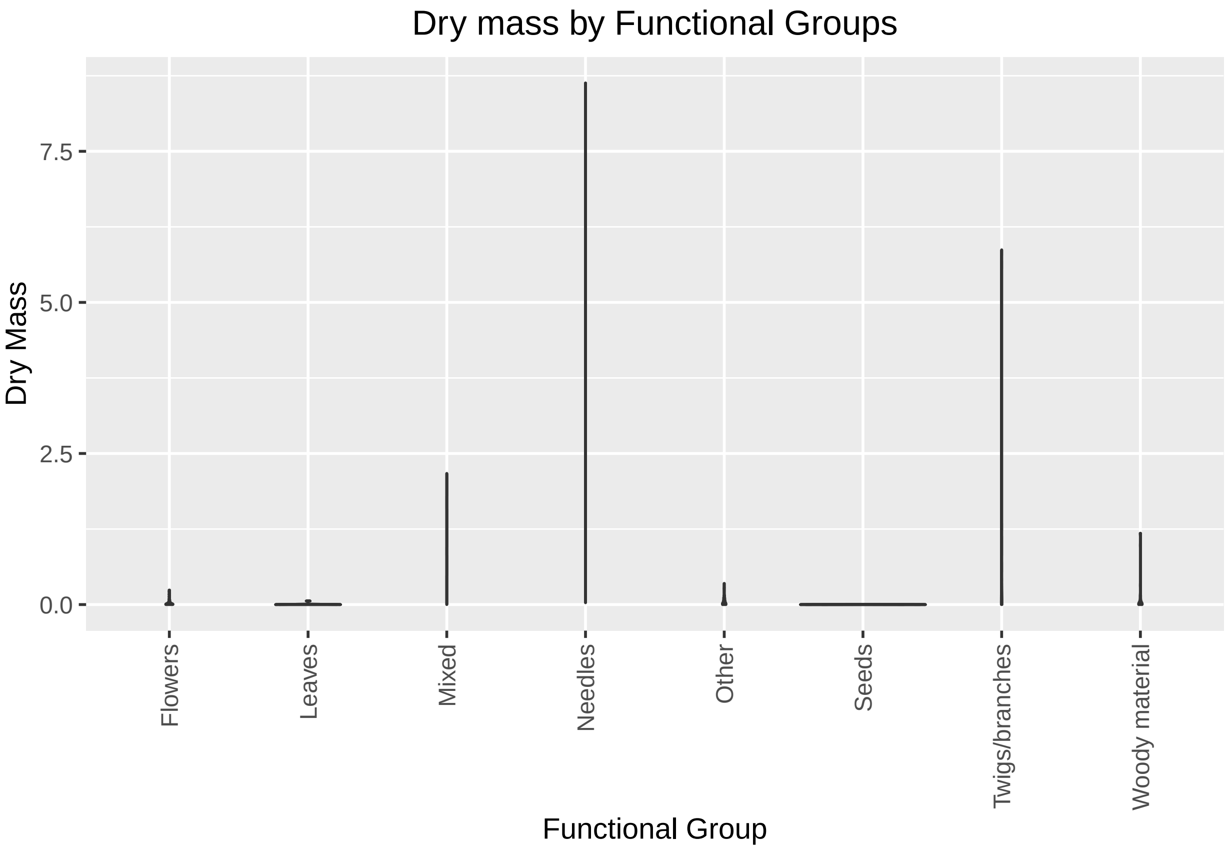
```
## Warning: Use of `Litter$dryMass` is discouraged.
## i Use `dryMass` instead.
```



```r
violin_drymass<-ggplot(data=Litter)+
  geom_violin(aes(x=Litter$functionalGroup,y=Litter$dryMass),show.legend=TRUE,draw_quantiles=c(0.25, 0.5
  theme(legend.position="bottom",axis.text.x=element_text(angle=90,vjust=0.5,hjust=1),plot.title=element
  labs(title="Dry mass by Functional Groups",x="Functional Group",y="Dry Mass")
violin_drymass
```

```
## Warning: Use of `Litter$functionalGroup` is discouraged.
## i Use `functionalGroup` instead.
```

```
## Warning: Use of `Litter$dryMass` is discouraged.
## i Use `dryMass` instead.
```



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

> Answer: We are interested to know how the dry mass compares to different functional groups, and the bar plot could show us the information with the range, percentile, mean, and outliers of the dry mass, while the violin plot can only prodivde an overview of the distribution of the dry mass. Thus, the box plot is a better visualization option in this case.

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

> Answer: Needles tends to have the highest biomass aross 21 terrestrial sites recorded in the NEON dataset.