Assignment 3: Data Exploration

Nicole Gutkowski

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.

```
library(tidyverse)
library(lubridate)
library(here)

getwd()
```

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

```
setwd(here())
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.

   Answer: A study of the ecotoxicology of neonicotinoids would be helpful in assessing the overall impact of these insecticides in agricultural operations. While we would want the neonicotinoids to work on insects harmful to crops, they need to remain safe for beneficial insect species such as pollinators.

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: Litter and woody debris can play roles in carbon and nutrient cycling in terrestrial and aquatic ecosystems. Additionally, litter and debris can provide habitats for organisms in the forest.

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. Mass data are measured separately for different types of litter and debris with an accuracy of 0.01 grams. 2.Sampling occurs at sites that have woody vegetation above 2 meters tall. 3. Ground traps are sampled once per year, and elevated trap sampling varies between biweekly, month, and bimonthly.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

Answer: The dimensions of the dataset are 4623 rows and 30 columns.

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...]

```
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
```

Answer: The most common effects studied are population, mortality, behavior, and feeding behavior. These effects are the most useful in determining the impacts of certain insecticides on removing harmful pests from agricultural operations, while not impacting beneficial insects.

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...]

```
summary(Neonics$Species.Common.Name, maxsum = 7)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee      Italian Honeybee
##                   152                   140                   113
##               (Other)
##                  3083
```

Answer: The six most commonly studied species in the dataset are the Honey Bee, Parasistic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. These species are all pollinators. They are of interest over other insects due to their essential nature in plant reproduction and growth.

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"
```

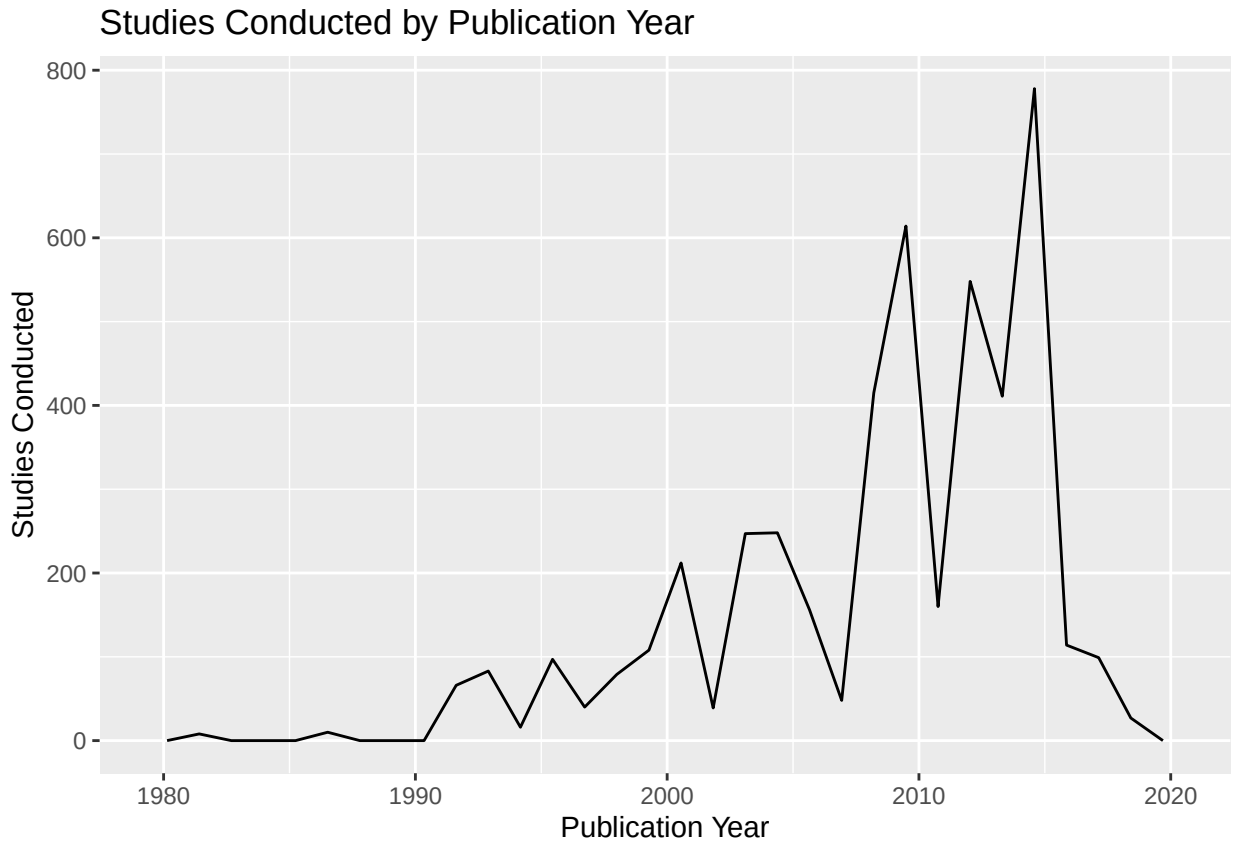
Answer: The class of `Conc.1..Author.` is a factor. This is because some of the data in the `Conc.1..Author.` column have characters such as , ~, <, >, or NR.

Explore your data graphically (Neonics)

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

```
ggplot(Neonics)+
  geom_freqpoly(aes(x = Publication.Year))+
  labs( title = "Studies Conducted by Publication Year",
        x = "Publication Year",
        y = "Studies Conducted")
```

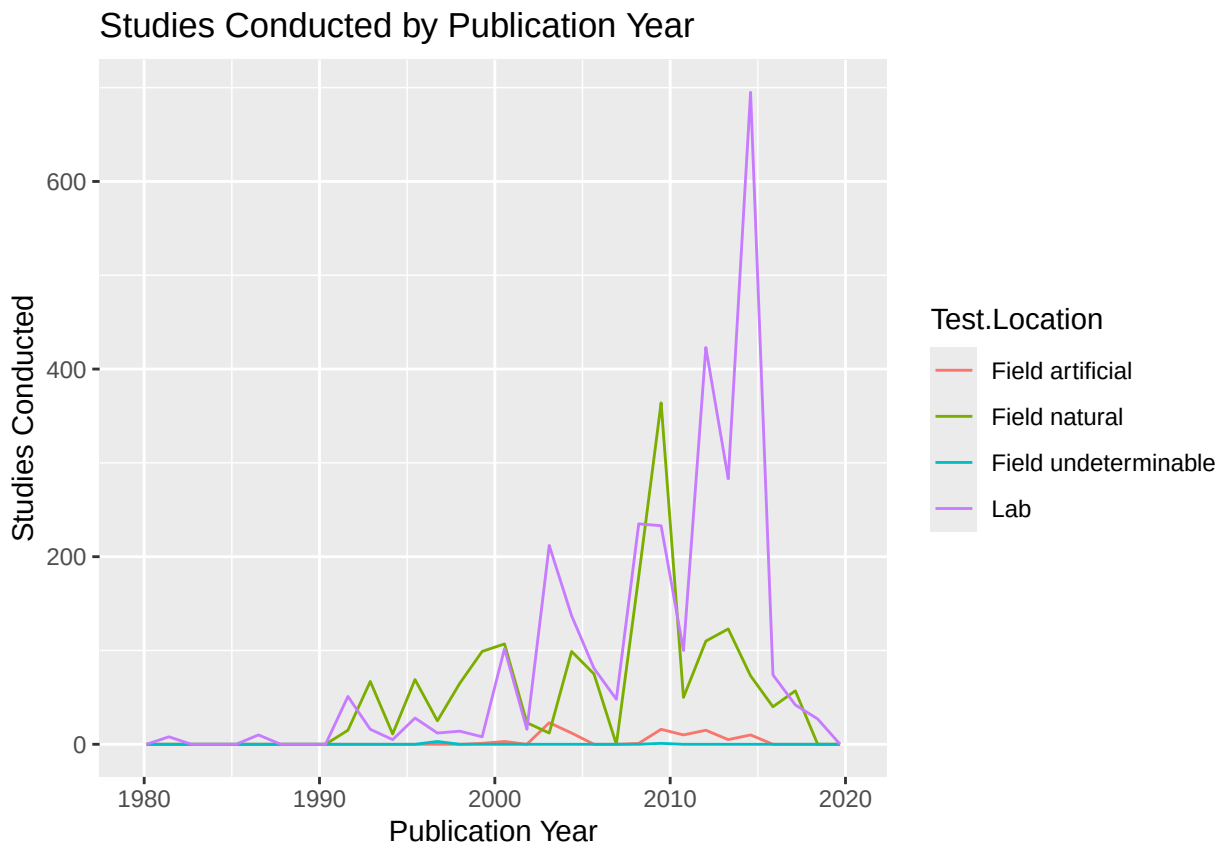
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
ggplot(Neonics)+
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location))+
  labs( title = "Studies Conducted by Publication Year",
        x = "Publication Year",
        y = "Studies Conducted")
```

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



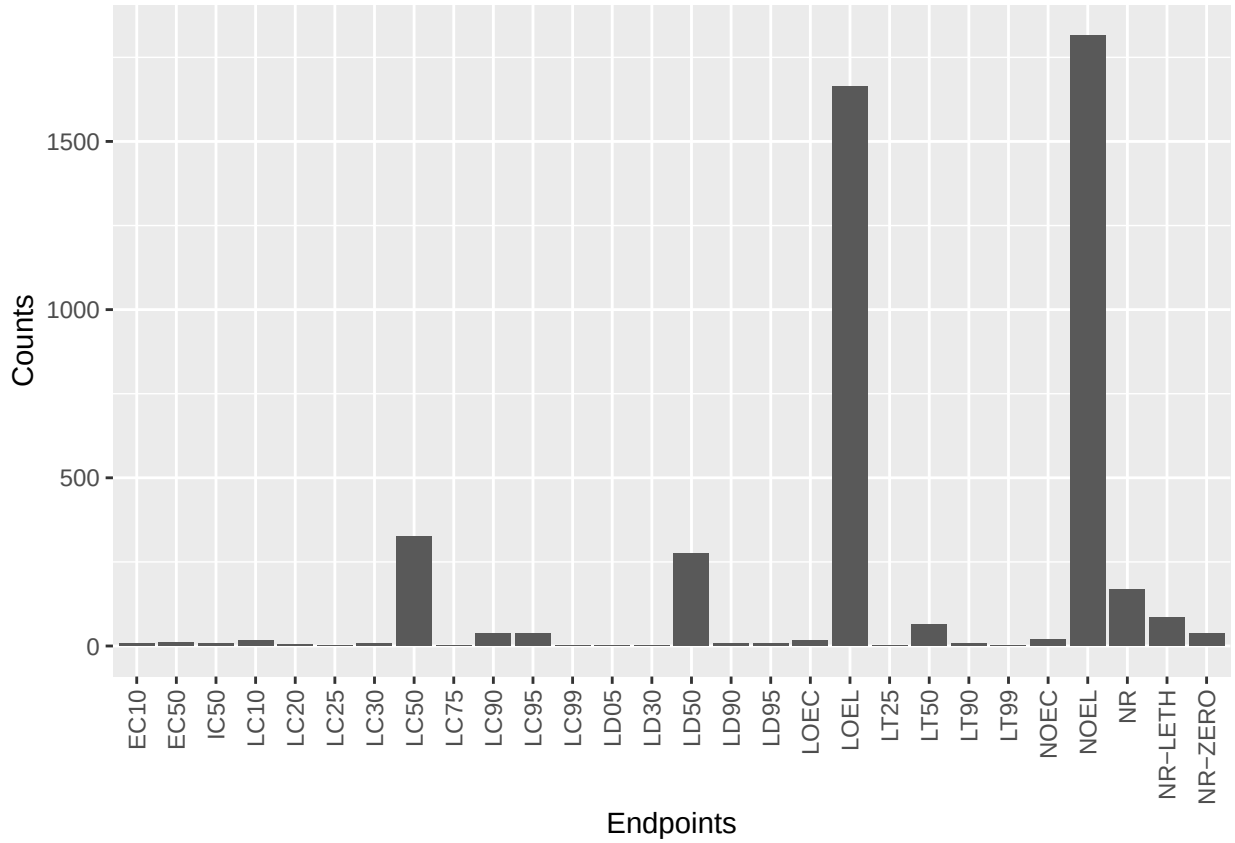
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer:The most common test locations are the lab and natural field sites. The testing at natural field sites generally increased between 1990 until 2010, and declined after 2010. The testing within the lab generally increased between 1990 and 2015, and declined after 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)+
  geom_bar(aes(x = Endpoint))+
  labs( x = "Endpoints",
        y = "Counts")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: The two most common endpoints are NOEL and LOEL. NOEL is the highest concentration that produces effects not significantly different than responses from the control. LOEL is the lowest concentration producing statistically significant effects.

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"
```

```
Date_Collection <- ymd(Litter$collectDate)
class(Date_Collection)
```

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

```
unique(Date_Collection)
```

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

Answer: The initial class of collectDate was a factor. When changed to a date, the new class is a date. Litter was sampled on August 2nd and 30th in 2018.

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`?

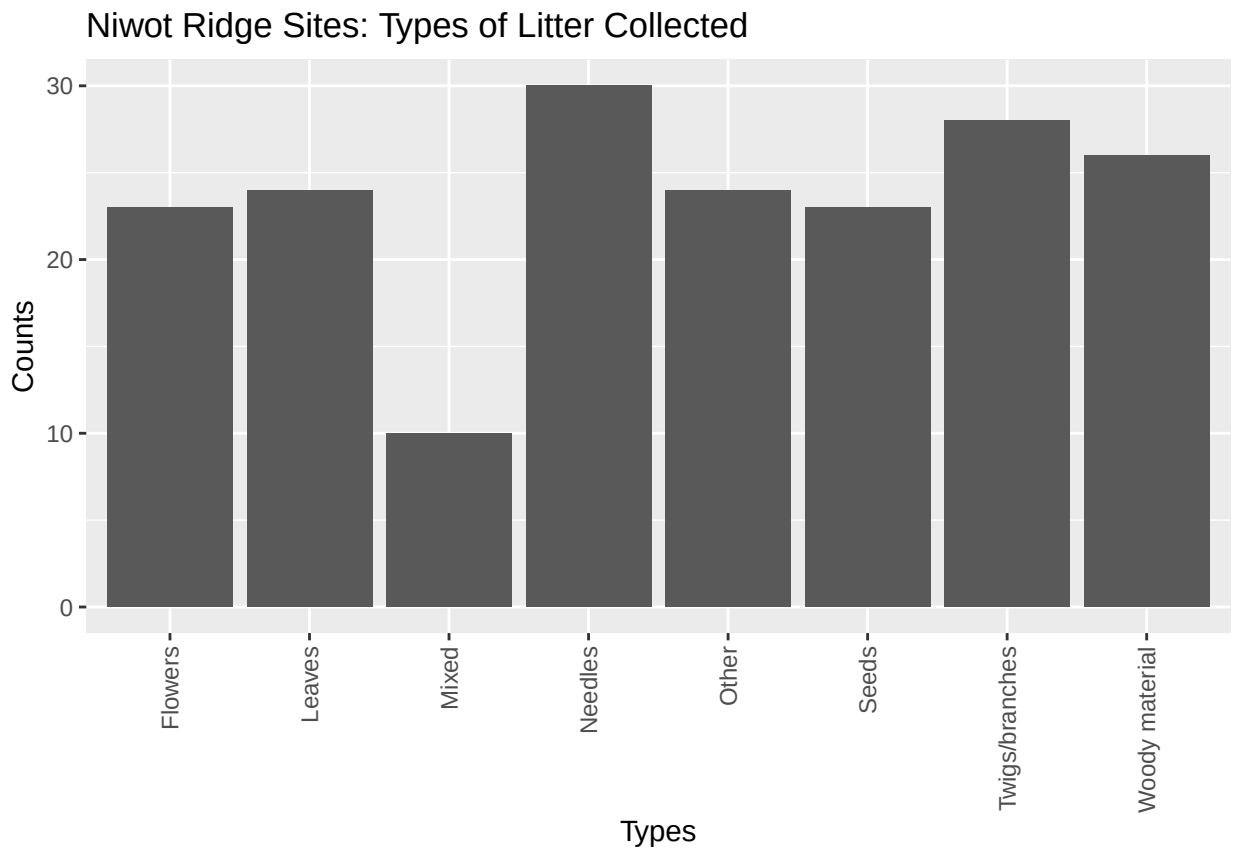
```
dif_plots <- unique(Litter$plotID)
summary(Litter$plotID)
```

```
## 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
```

Answer: 12 different plots were sampled at Niwot Ridge. The command `unique` provides you with the unique values under the indicated column. The command `summary` provides you with both the uniqe values and how many times that they occur.

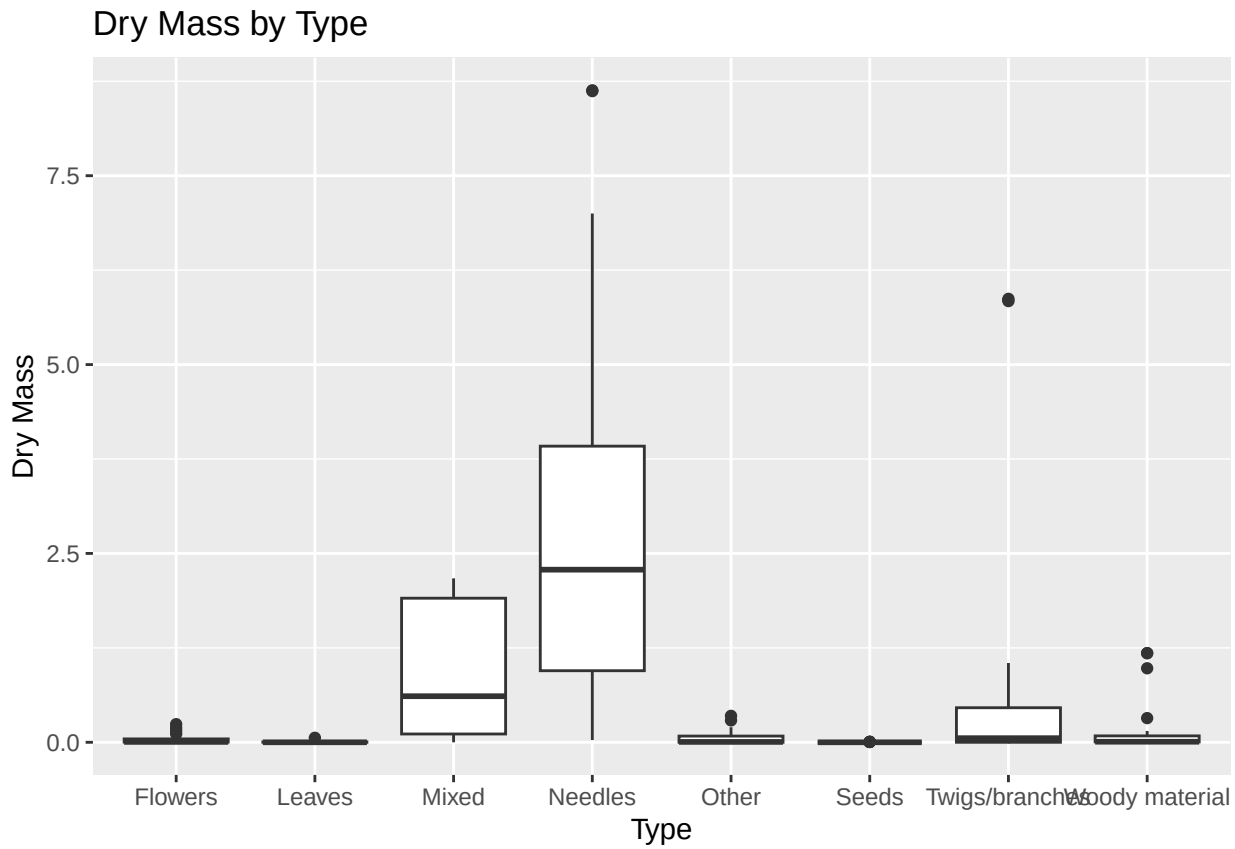
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter)+
  geom_bar(aes(x = functionalGroup))+
  labs( x = "Types",
        y = "Counts",
        title = "Niwot Ridge Sites: Types of Litter Collected")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

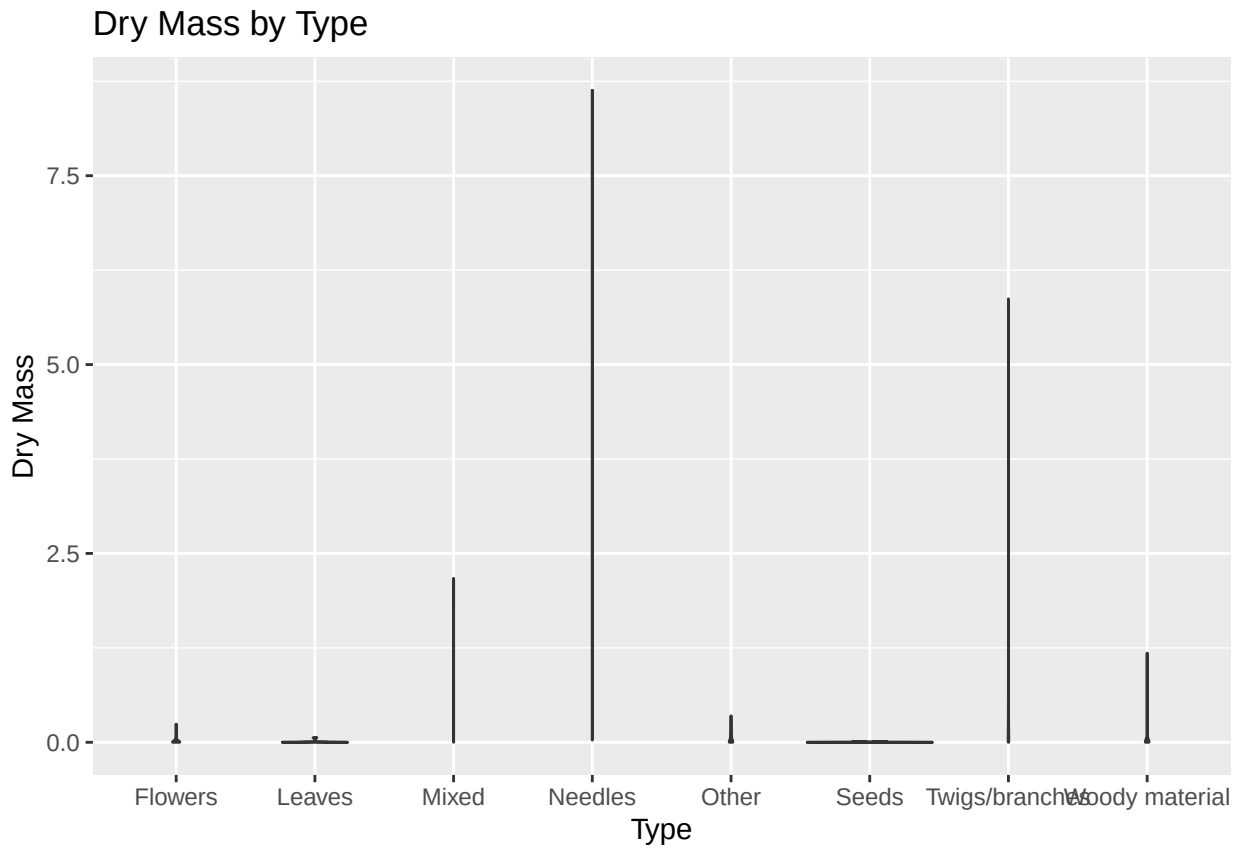


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(Litter) +
  geom_boxplot(aes(functionalGroup, dryMass))+
  labs( x = "Type",
        y = "Dry Mass",
        title = "Dry Mass by Type")
```




```
#
ggplot(Litter) +
  geom_violin(aes(functionalGroup, dryMass))+
  labs( x = "Type",
        y = "Dry Mass",
        title = "Dry Mass by Type")
```



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

> Answer: A violin plot is most effective at showing the distribution and density of the data. However, because most of our data is clustered around very light masses, a boxplot is more effective in our case. Boxplots can provide a clear visualization of central tendency, without relying on data density.

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

> Answer: The types of litter that have the highest biomass are mixed, needles, and twigs/ brances.