Assignment 3: Data Exploration

Maeve Arthur

Spring 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.

```
#checking working directory
getwd()
```

```
## [1] "/home/guest/EDA-Spring2023"
```

```
#loading necessary packages
library(tidyverse)
library(lubridate)

#reading two data sets
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",stringsAsFactors = TRUE)

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

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: We might be interested in the ecotoxicology of neonicotinoids on insects because of the effects that spraying the insecticides has on humans. We might also be interested in learning about the effects of the iinsecticide on insect biodiversity and abundance, and how that effects overall ecoosystem health.

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: We miight be intersted in studying litter and woody debris that falls to the ground in forests because of its potential to start and spread forest fires. Analyzing this data may also tell about rates of forest decay.

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. Sampling of litter and woody debris only occurs in sites with woody vegetation greater than 2m tall. 2. 1 to 4 litter traps are deployed in tower plot sampling sites. 3. Traps are sampled once every year.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
# using dim() functiont to find dimensions of dataset
dim(Neonics)
```

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

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?

```
# Using summary() on Neonics 'Effect' column
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 effect that is studied is population. The second moost common effect is mortality. These miight be of special interest because researches likely most want to know how well the insecticide works at killing 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: The `sort()` command can sort the output of the summary command...]

```
# summarizinig species common name and then viewing it
common_names <- summary(Neonics$Species.Common.Name)
# view(common_names)
```

Answer: The most commonly studied insect species are: Honey Bee, Parasitic Wasp, Buff Tailed Bumble Bee, Carniolan Honey Bee, Bumble Bee, Italian Honeybee. These are likely the most commonly studied because of their importance as pollinators.

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?

```
# using class function
class(Neonics$Conc.1..Author.)
```

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

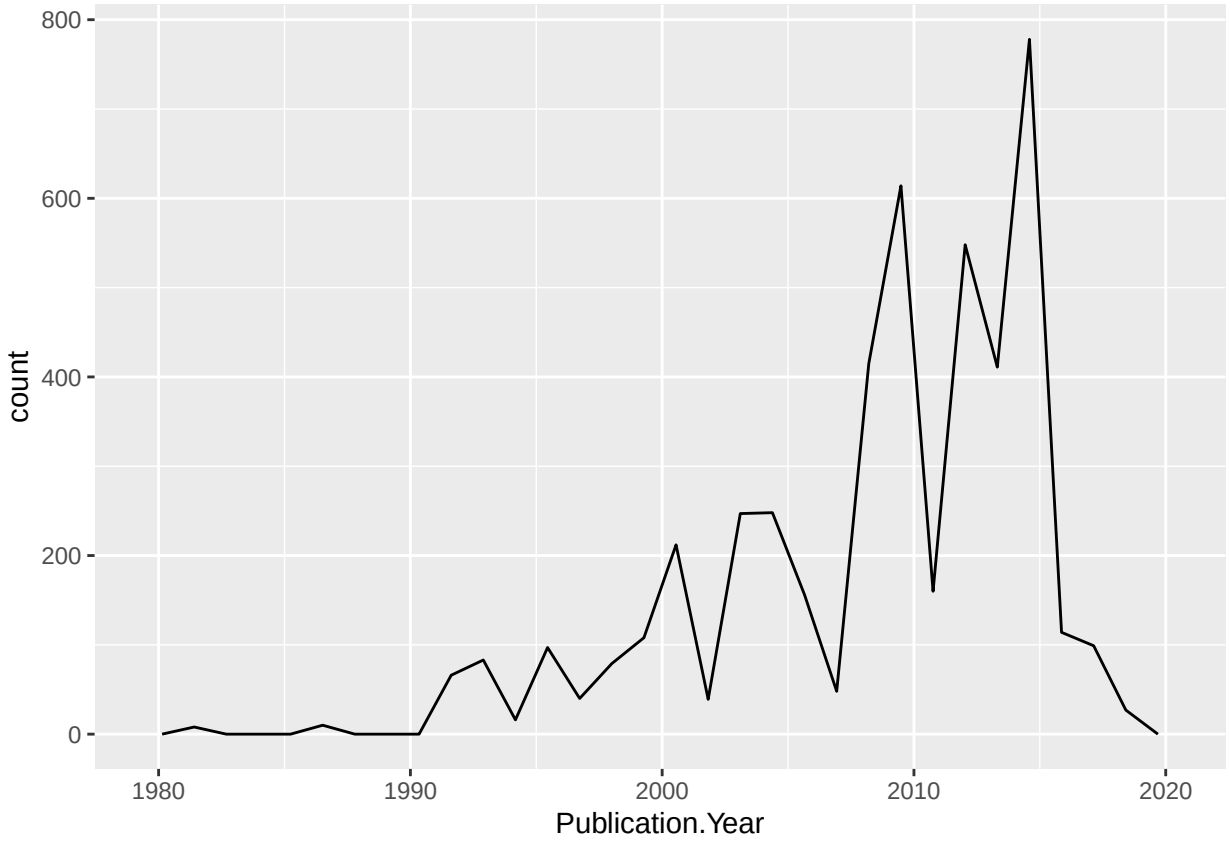
Answer: It is a factor. It is not a numeric class because it is categorical data.

Explore your data graphically (Neonics)

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

```
# Plotting number of studies conducted by publication year
ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year))
```

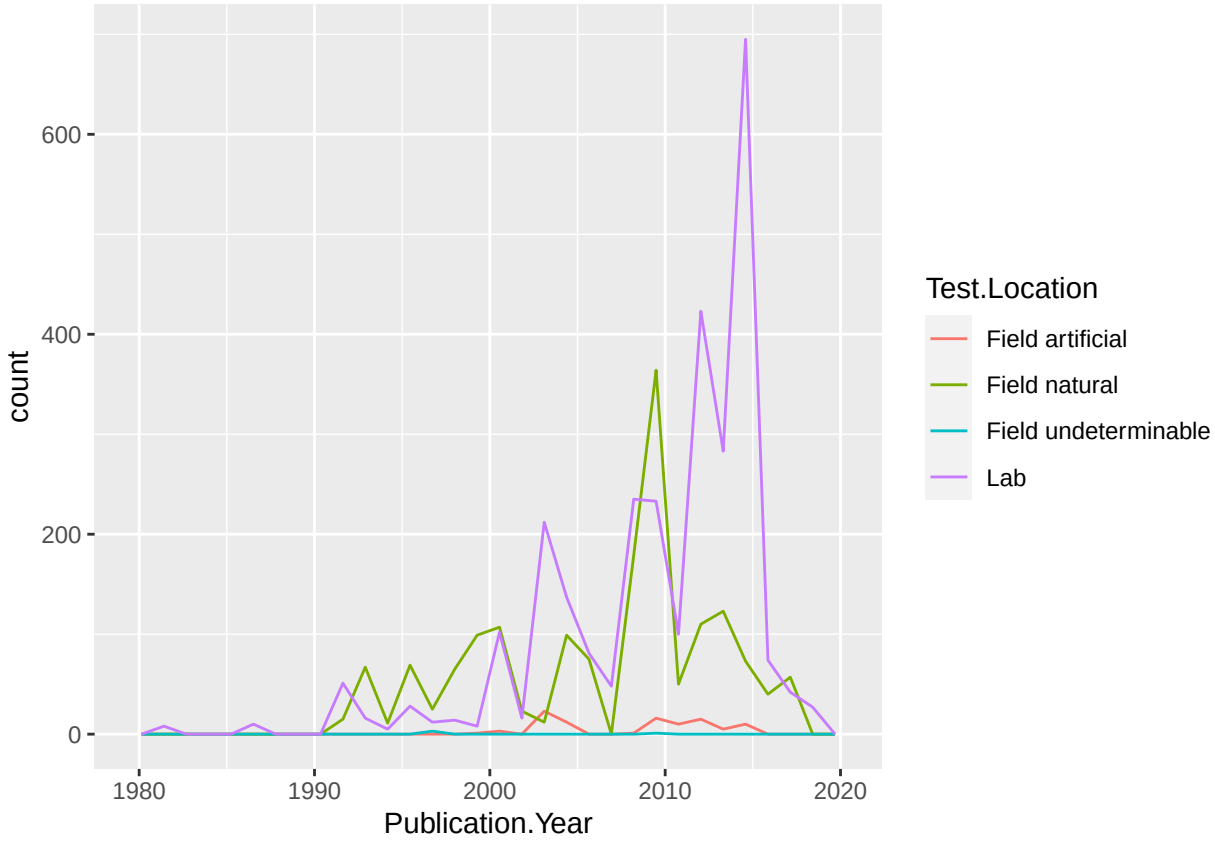
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
# adding a color aesthetic so that different Test.Location are displayed as
# different colors
ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year, color = Test.Location))
```

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



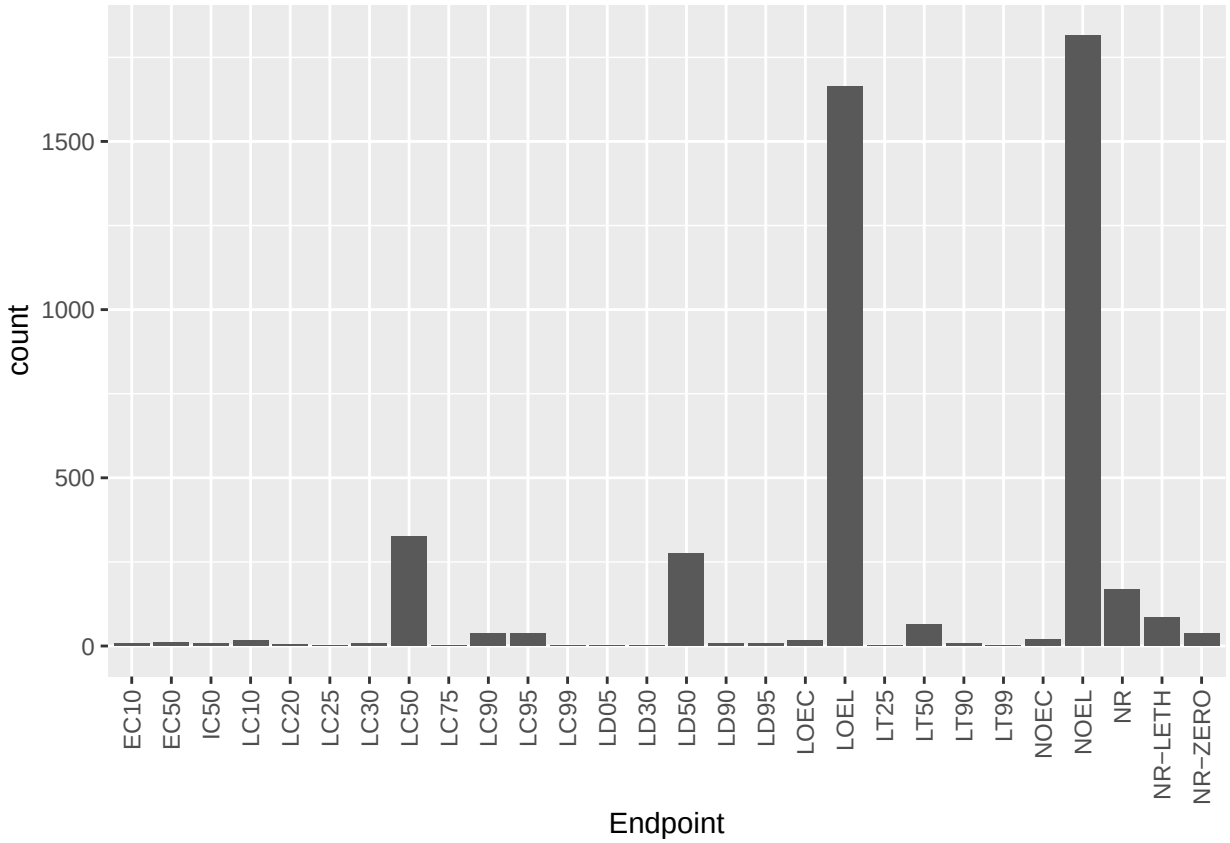
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test location appears to be the lab. Around 2015, the frequency of lab testing locations spiked, then they fell again. The second most common testinig location is field natural. This increasd in frequency just before 2010, then fell again. The other testing methoods, field artifical and field undeterminable, are infrequent throughout the time period.

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...]

```
# Using ggplot and geombar to plot Endpoint counts
ggplot(Neonics, aes(x = Endpoint)) + geom_bar() + theme(axis.text.x = element_text(angle = 90,
    vjust = 0.5, hjust = 1))
```



Answer: The two most common end points are LOEL and NOEL. LOEL is defined as "Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls." NOEL is defined as "No-observable-effect-level: highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test."

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.

```
# determining class of collectDate
class(Litter$collectDate)
```

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

```
# changing collectDate to a date
as.Date(Litter$collectDate)
```

```
##   [1] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##   [6] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [11] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [16] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [21] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [26] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [31] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [36] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [41] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
```

```
##  [46] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [51] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [56] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [61] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [66] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [71] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [76] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [81] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [86] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [91] "2018-08-02" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
##  [96] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [101] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [106] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [111] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [116] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [121] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [126] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [131] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [136] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [141] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [146] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [151] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [156] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [161] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [166] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [171] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [176] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [181] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [186] "2018-08-30" "2018-08-30" "2018-08-30"
```

```
view(Litter$collectDate)

# Determining which dates litter was sampled in August 2018
unique(Litter$collectDate)
```

```
## [1] 2018-08-02 2018-08-30
## Levels: 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`?

```
# unique function tells us how many plots were sampled at Niwot Ridge
length(unique(Litter$plotID))
```

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

```
# contrasting unique function with summary function
summary(Litter$plotID)
```

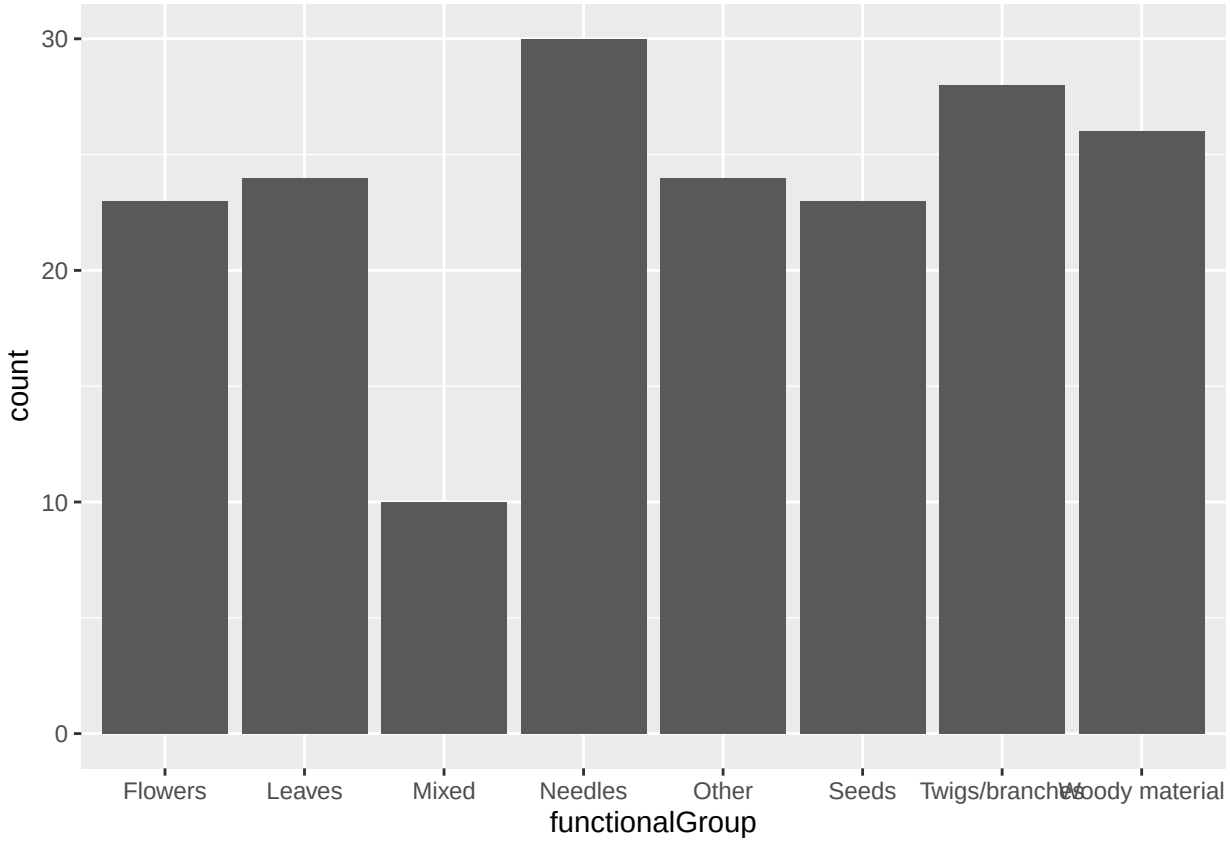
```
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061
##       20       19       18       15       14        8       16       17
## NIWO_062 NIWO_063 NIWO_064 NIWO_067
##       14       14       16       17
```

Answer: 12 plots werre sampled at Niwot Ridge. The informatioin obstained from 'unique' eliminates duplicate values in a vector, returning what is unique. Summary, on the other hand, returns a statistic about the vector. In this case, when I ran the summary function on siteID, it

gives me the count of how many plots were sampled at Niwot Rdige.

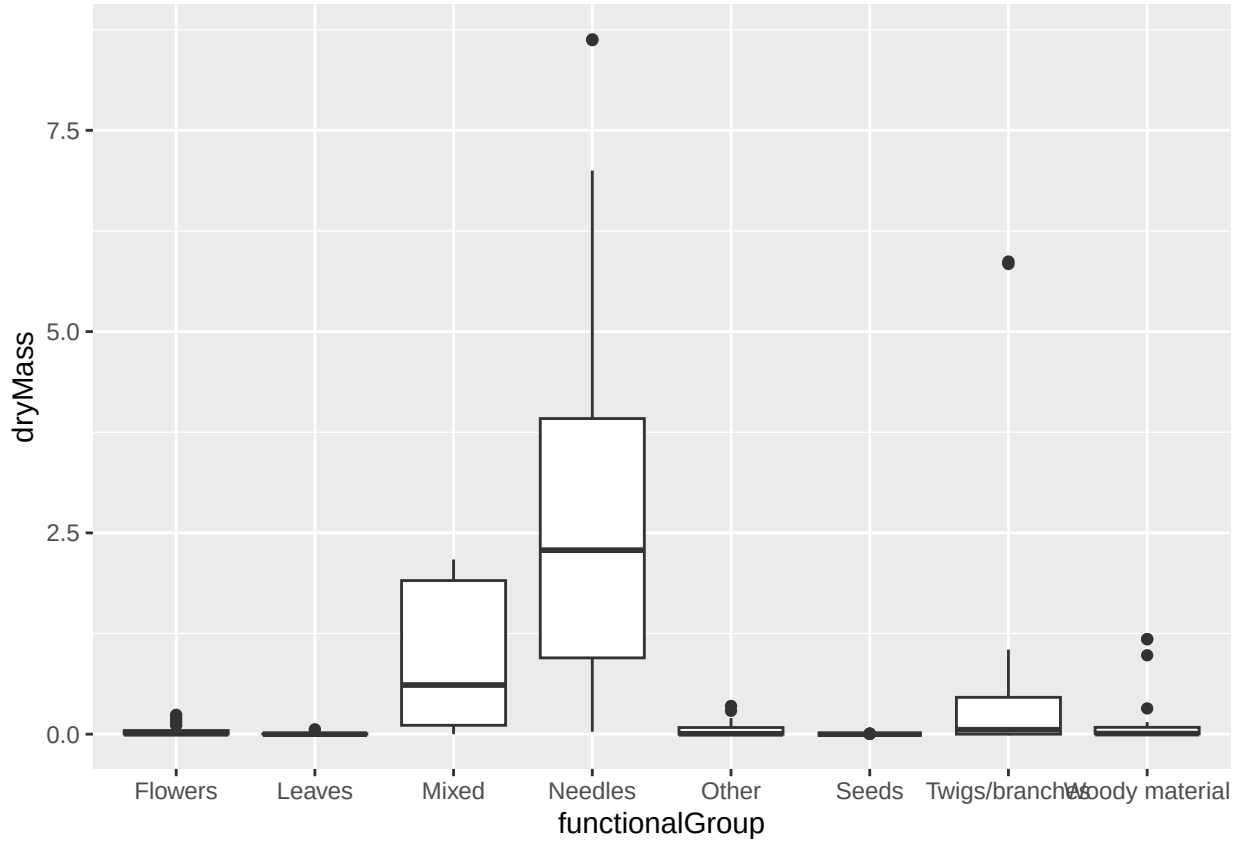
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
# Using ggplot and geombar to make a bar graph
ggplot(Litter, aes(x = functionalGroup)) + geom_bar()
```

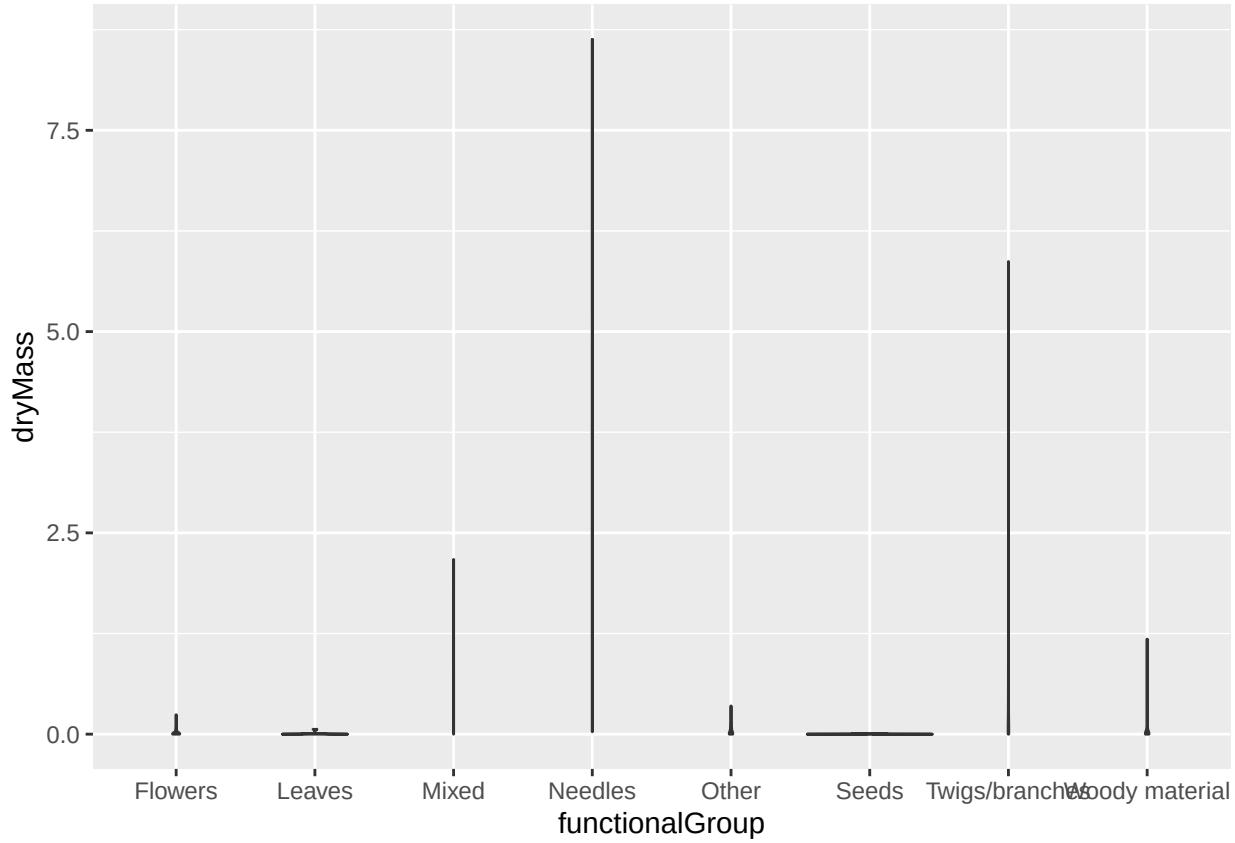


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
# Boxplot
ggplot(Litter) + geom_boxplot(aes(x = functionalGroup, y = dryMass))
```




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



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

> Answer: The boxplot is more effective than the violin plot in this case because it shows the range of diistribution and gives the summary statistics for dryMass, which is a numerical variable. The violin plot here reveals very little information about the data because the violin plot uses density curves, which is not useful for this data type.

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

> Answer: The litter types with the highest biomass are needles and mixed. Twigs and brances appears to have the 3rd highest biomass.