Assignment 3: Data Exploration

Camila Rodriguez

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.

```r
#install packages and add library function below; add relative paths
#install.packages('tidyverse')
#install.packages('lubridate')
#install.packages('here')

library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
library(here)
```

```
## here() starts at /home/guest/EDE_Fall2024
```

```r
Neonics<- read.csv('Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv')
Litter<- read.csv('Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv');
stringsAsFactors = TRUE

#print(Litter)
#print(Neonics)
```

##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: Application of Neonicotinoids can have unintentional consequences on non-target inseects, such as pollinators and other beneficial insects. This insectiside is commonly used in agriculture for crop rotation farming, and poses many threats to ecological diversity. Plants can also be contaminated with this insectiside, making it toxic and sub-toxic for moths and butterflies that consume these plants.

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: Available biomass (in the form of litter/woody debris) from the understory or ovrgrowth of vegetation can be an indicator of high flammabiltiy given the right conditions for a fire (heat, drought,etc.) Soil, weather conditions and available biomass are all factors that are considered in fire risk calculations.

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 fine woody debris sampling is sampled from terrestrial NEON sites that contain woody vegetation greater than 2m in height. 2. Sampling only occurs in tower plots, and there is one litter trap pair (one elevated trap and one ground trap) used for every 400 m2 plot area, resulting in 1-4 trap pairs per plot. 3. Trap placement within plots may be either placed in targeted or randomized locations, depending on the vegetation (if >50% of land is covered in vegetation, traps are placed randomly; is <50% is covered, then traps are only placed under qualifying vegetation)

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset? > Answer: 4623 rows, 30 columns

```
nrow(Neonics)
```

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

```
ncol(Neonics)
```

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

```
dim(Neonics)
```

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

```
#both functions give dimensions
```

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)

sorted_summary<- summary(Neonics$Effect)
sort(sorted_summary)
```

   Answer:The most common effects are avoidance, behavior, enzymes, development, feeding behavior and genetics. Pesticides can affect not only the neurosystem but can also alter an insect's behavior, change its genetic sequences and impact behavior/feeding patterns.

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$Species.Common.Name)
sort(Neonics$Species.Common.Name)
#View(Neonics$Species.Common.Name)
```

   Answer: The 6 most commonly studied species are: Honey Bee, Buff Tailed Bumblebee, Parasitic Wasp, Carniolan Honey Bee, Italian Honeybee, Asian Lady Beetle

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…]

```
#View(Neonics)
```

   Answer: Some values are NR, which RStudio does not recognize as a numeric value but as a charectar value

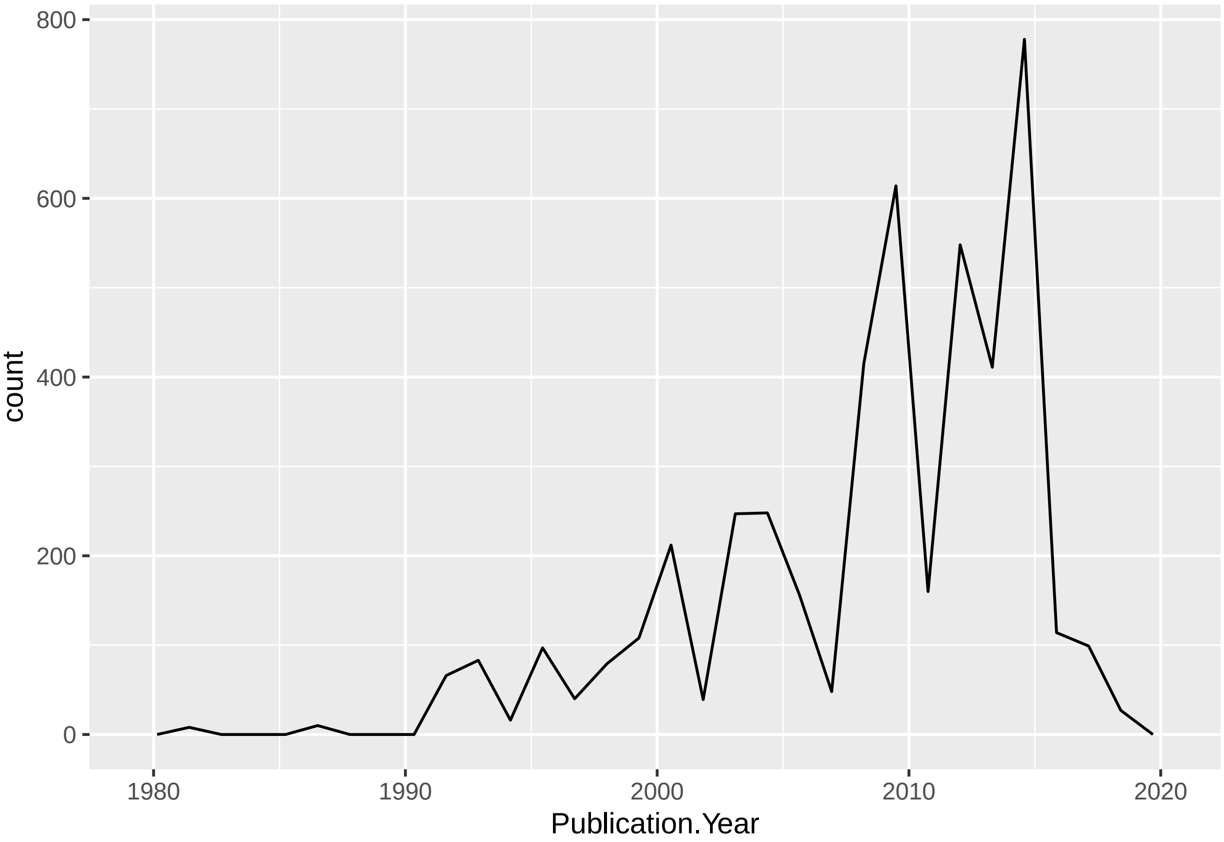
Explore your data graphically (Neonics)

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

```
ggplot(Neonics)+

geom_freqpoly(aes(x=Publication.Year))
```

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

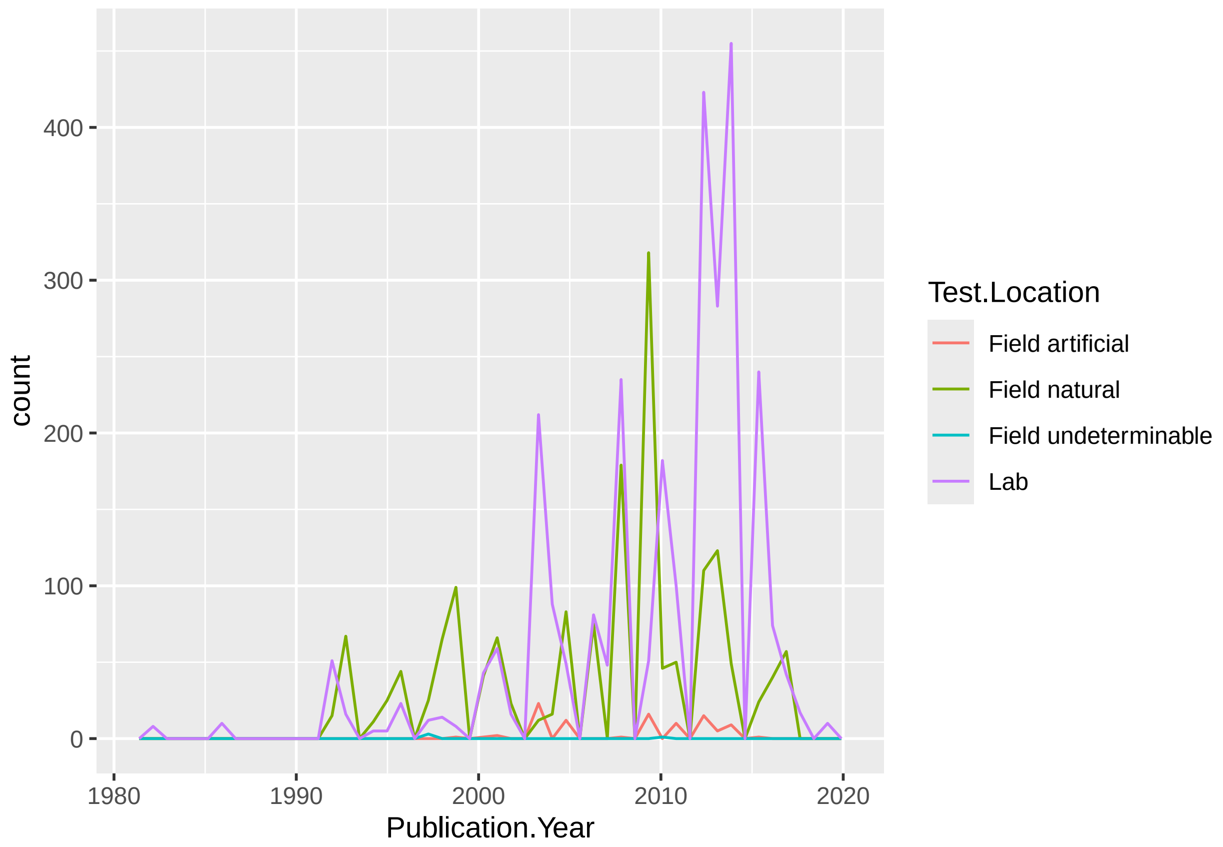


```
#first need to add ggplot of dataset, then specify type of plot and any
```

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), bins=50)
```



```
#column can be selected as a type of color to represent publication year by
```

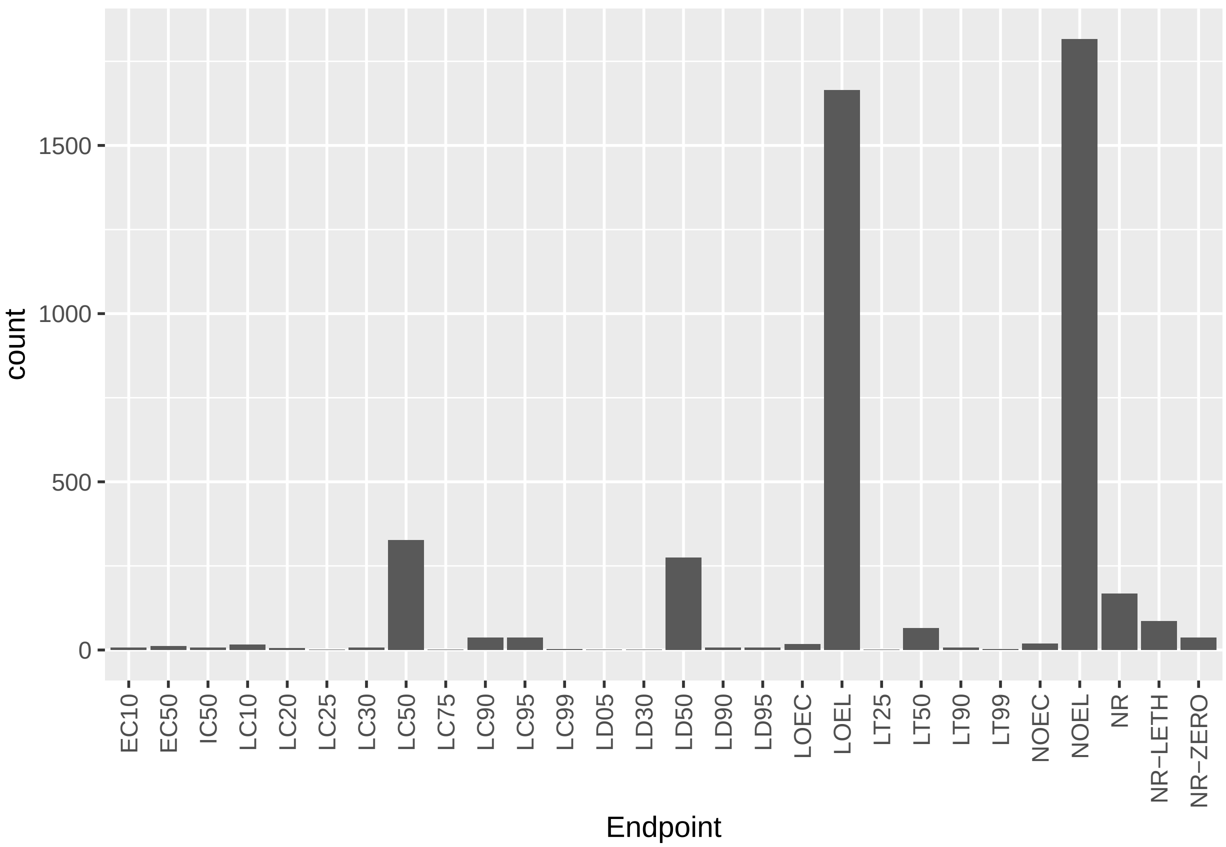
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The lab is the most common test location. The number of tests taken at labs drastically increases between 2010-2020.

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…]

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



Answer:LOEL and NOEL are the most common endpoints. LOEL (Lowest-observable -effect-level) indicates that even at the lowest dose (concentration), there produced effects are significantly different from responses of controls. NOEL (No-observable-effect-level) is defined as not having any significant difference in effects even with the highest dosage compared to controls.

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.

```
#verify class, if not 'date' then use as.Date function to change it from
class(Litter$collectDate)
```

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

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

#use the unique function to filter unique dates within a specified time range


august_samples <- unique(Litter$collectDate[Litter$collectDate >= "2018-08-01"

august_samples
```

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

```
#call back the command to print out unique dates from August 2018
#specify dataset and column within data set first, then specify date range after
```

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`?

```
#summary(Litter$plotID)

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

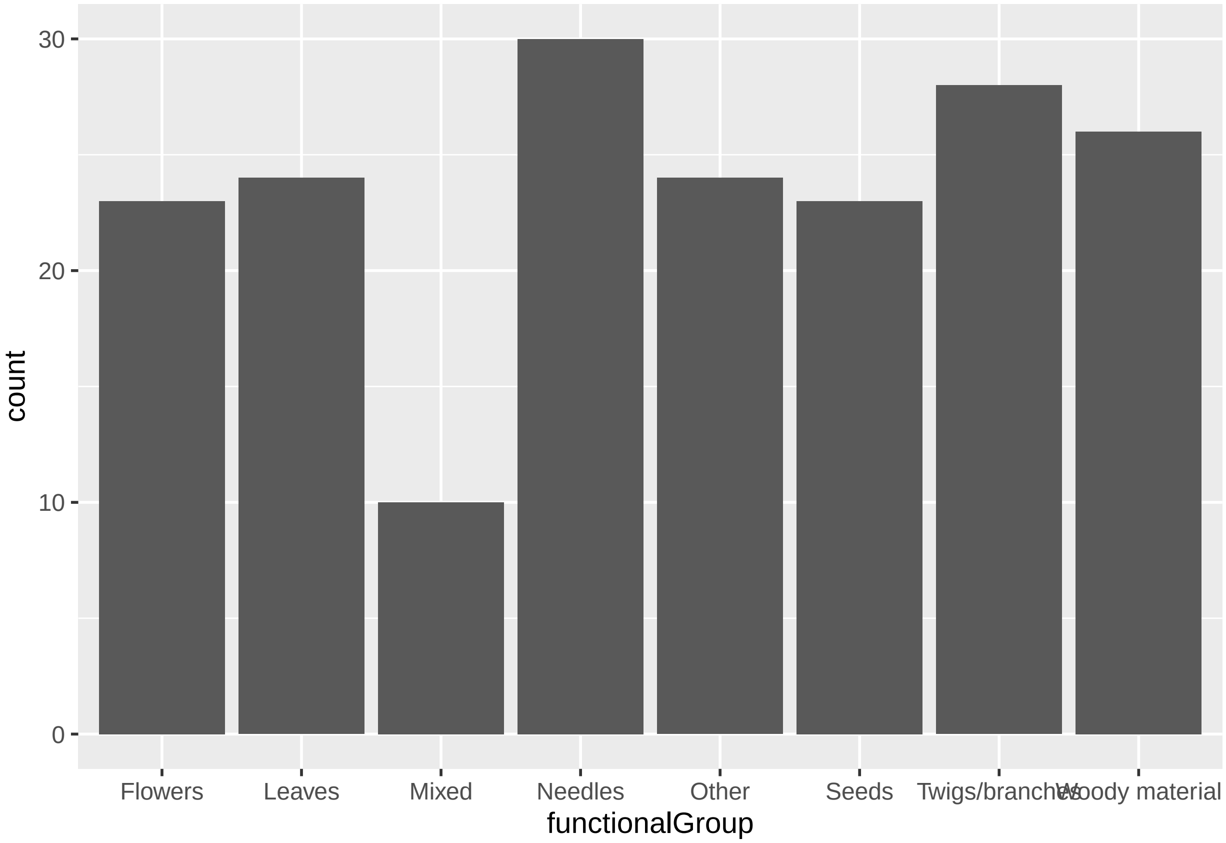
```
##  [1] "NIWO_061" "NIWO_064" "NIWO_067" "NIWO_040" "NIWO_041" "NIWO_063"
##  [7] "NIWO_047" "NIWO_051" "NIWO_058" "NIWO_046" "NIWO_062" "NIWO_057"
```

```
#unique_values_col prints back unique values after you specify a column
```

Answer: There are 12 unique plot values, however since there are 188 total entries shown in the summary, multiple samples were taken across the 12 plots.

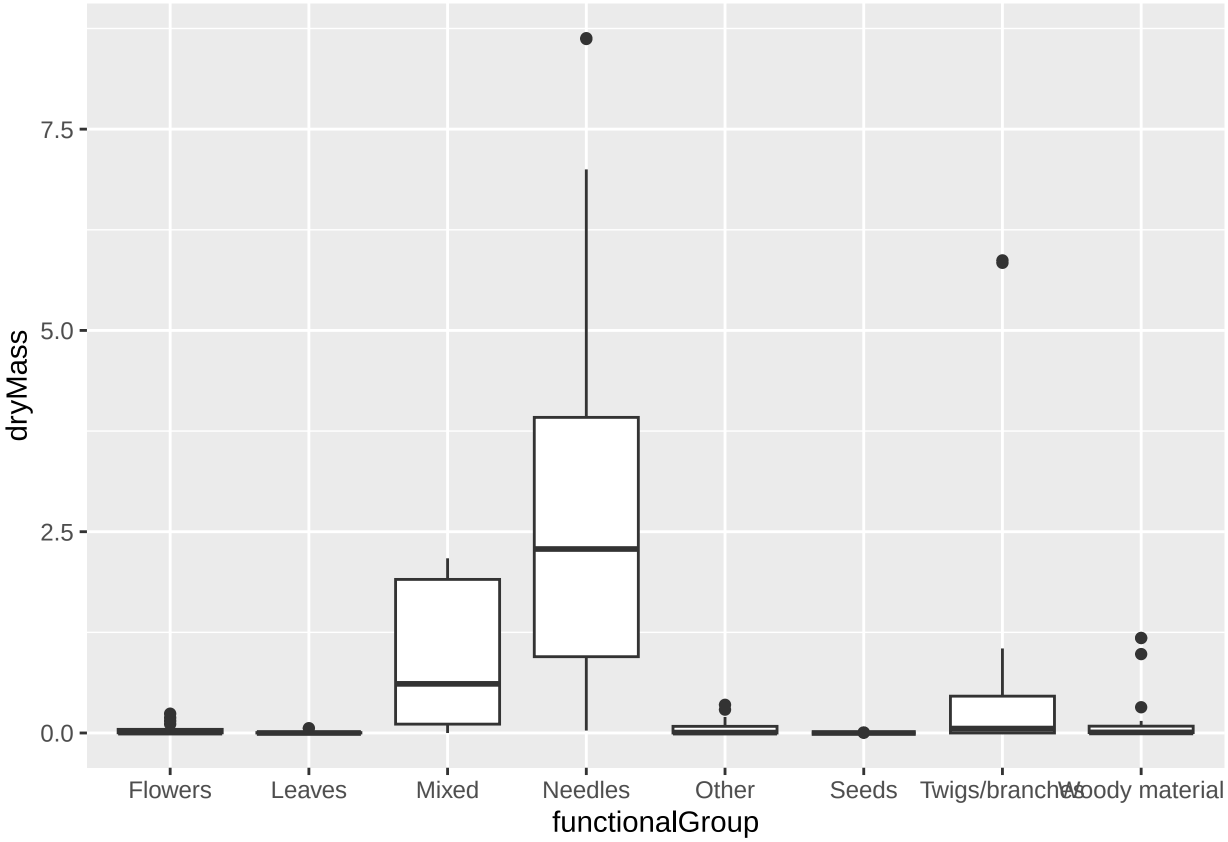
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(data=Litter, aes(x=functionalGroup))+
        geom_bar()
```

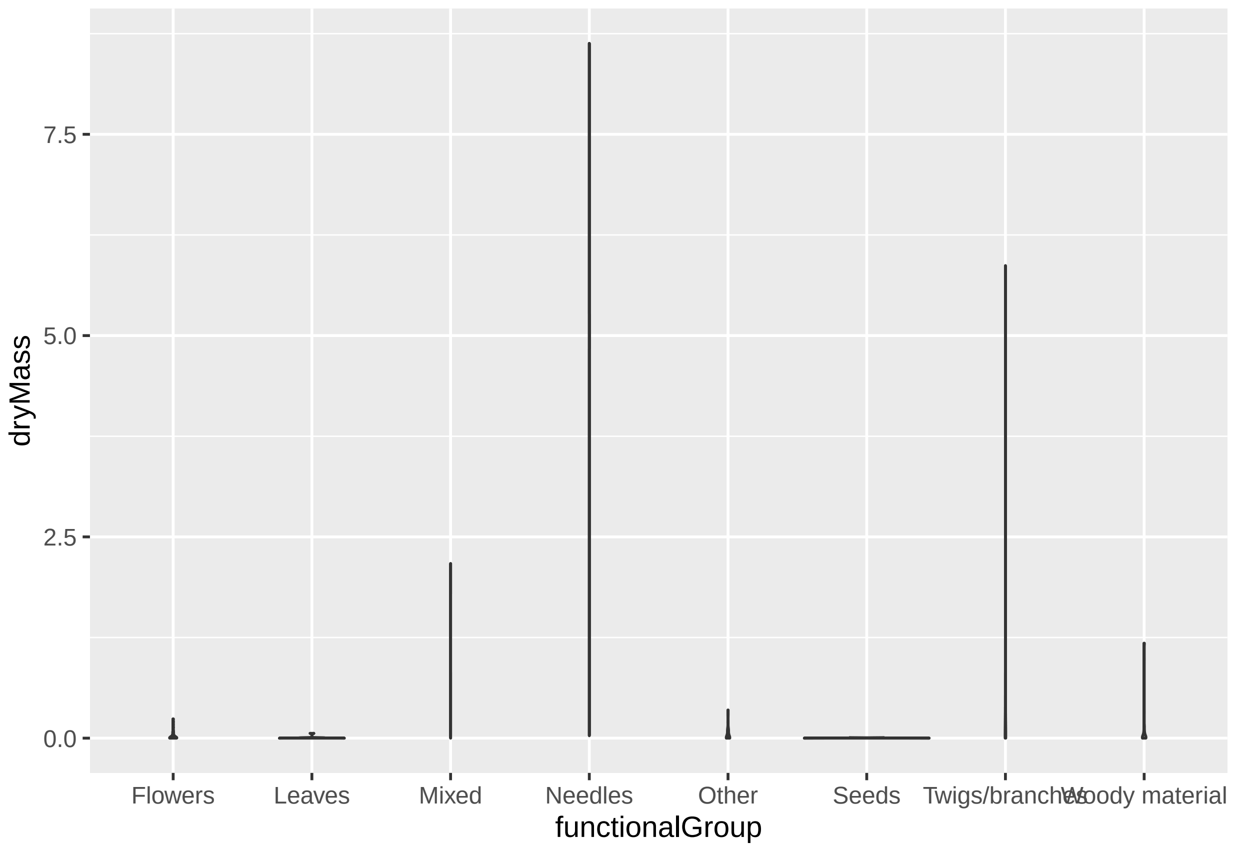


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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



```
ggplot(data=Litter, aes(y=dryMass, x=functionalGroup))+
        geom_violin()
```



```
# x and y axis can be flipped but for visual purposes x=functionalGroup
#first, specify data set with parameters then add the type of graph
```

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

> Answer: The boxplot is more effective in this case because it shows the distribution of biomass of different litter types, while the violin plot provides the max recorded value of biomass but it does not have much any width on the bars since there is not sufficient data in the different litter categories.

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

> Answer: Needles and mixed litter tend to have the highest bimass, which can be seen by the height of the line (range) and also by the white boxes in the plot, which indicate higher biomass than other litter types.