Assignment 3: Data Exploration

Camber Vincent

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.

```
#Load necessary packages
library(tidyverse)
library(lubridate)
library(here)

getwd() #Check current working directory
```

```
## [1] "/Users/cambervincent/EDA_Fall_2024/Assignments"
```

```r
#Upload the datasets, ensuring strings read as factors
Neonics<-read.csv(here("Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"),
                    stringsAsFactors = TRUE)
Litter<-read.csv(here("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: Studies on the ecotoxicology of neonicotinoids on insects are likely of interest due to the possibility for the insecticide's adverse impacts on non-target species. Neonicotinoids, like other environmental toxins, have a high risk of environmental impact through damage to non-target species. We likely study neonicotinoids in order to determine specific concentrations that cause damage, concentrations that can be applied without causing adverse impacts, and to characterize the general impacts of insecticide use.

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: Forest litter and woody debris are important in ecosystems for nutrient cycling, as the decomposition of ground floor debris will eventually replenish nutrients in the forest soil. Litter and debris may also provide important habitat for certain wildlife, contribute to soil stabilization, or serve as indicators of ecological health. We likely study the litter and debris in order to characterize the type, amount, seasonality, and other factors of the debris to better understand the ecosystem and its state of health.

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 executed at terrestrial NEON sites that contain woody vegetation >2m tall 2. Trap placement within plots may be either targeted or randomized, depending on the vegetation 3. Ground traps are sampled once per year

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
dim(Neonics) #Check dimensions
```

```
## [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? [Tip: The **sort()** command is useful for listing the values in order of magnitude…]

```
sort(summary(Neonics$Effect)) #Sort the summary of the Effect column
```

```
##       Hormone(s)        Histology       Physiology          Cell(s)
##                1                5                7                9
##     Biochemistry     Accumulation     Intoxication    Immunological
##               11               12               12               16
##       Morphology           Growth        Enzyme(s)         Genetics
##               22               38               62               82
##        Avoidance      Development     Reproduction Feeding behavior
##              102              136              197              255
##         Behavior        Mortality       Population
##              360             1493             1803
```

Answer: The most common effects studied are on population, mortality, behavior, feeding behavior, and reproduction. These are mostly likely the effects of interest as they correlate to how effective an insecticide would be. Most insecticides aim to have an adverse impact on population, increase mortality, decrease reproduction, or influence behavior adversely. Ecotoxicology studies can be carried out to determine the effects of neonicotinoids on target species to determine effectiveness or non-target species to determine adverse impacts of application.

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…]

```
#Sort the summary of the Species Common Name column
sort(summary(Neonics$Species.Common.Name,maxsum=7))
```

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

```
  #maxsum arg. set to 7 to represent six most common species and "other" category
  #maxsum auto set to 100 without specification
```

Answer: The six most commonly studied species in the dataset are the Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honey Bee respectively. Most of these species are pollinators, playing an important role in the ecosystem at large and in the agricultural sector specifically. Parasitic wasps also play a special role in biological pest management that may make them important. All of the species live in complex social structures, so the investigations into "behavior" may be of more importance for insects that live in hives than other types of insects. These species are likely of interest over other insects due to their ecological, and thus economic, importance, their declining populations, and conservation efforts.

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.) #Class of `Conc.1..Author1 column
```

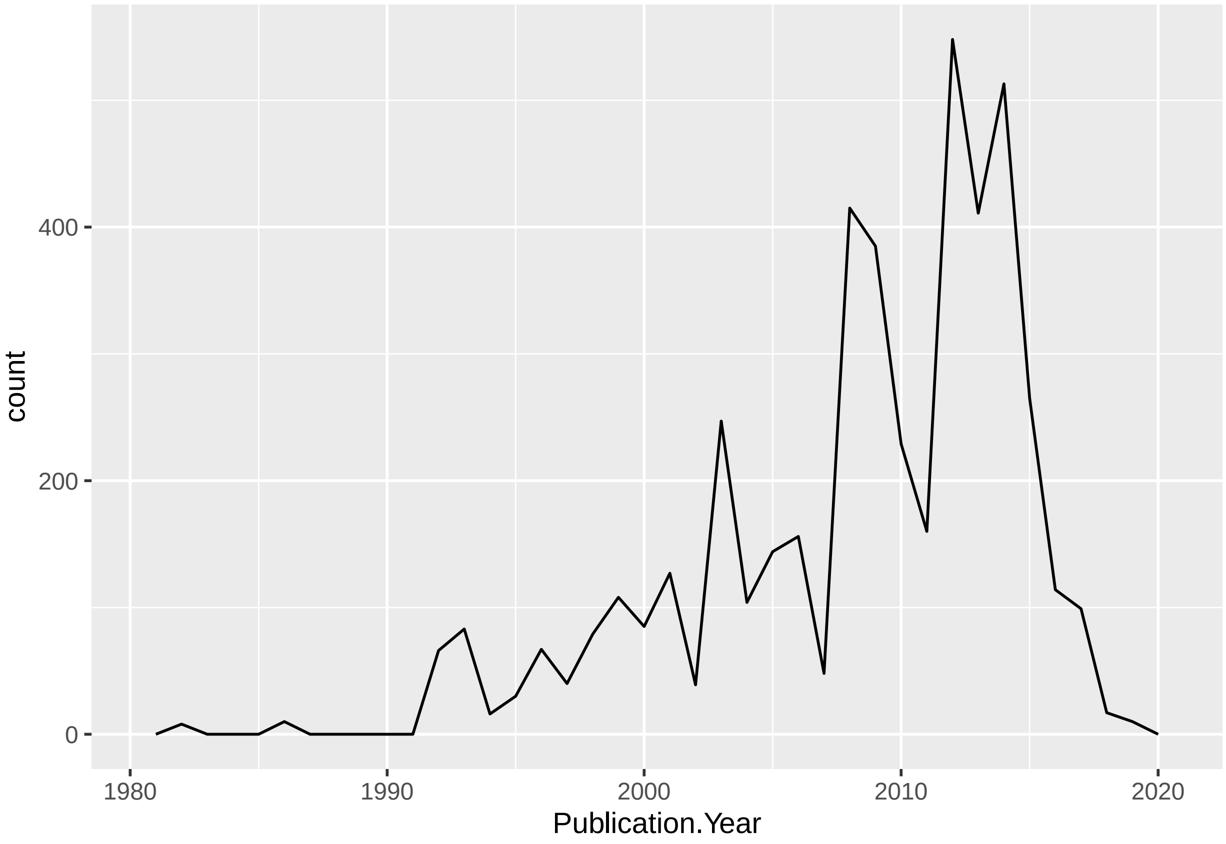
```
## [1] "factor"
```

Answer: The class is factor, not numeric. The data is stored as a factor rather than as numeric due to the presence of a trailing "/" character in several of the concentration values stored in the dataset.

Explore your data graphically (Neonics)

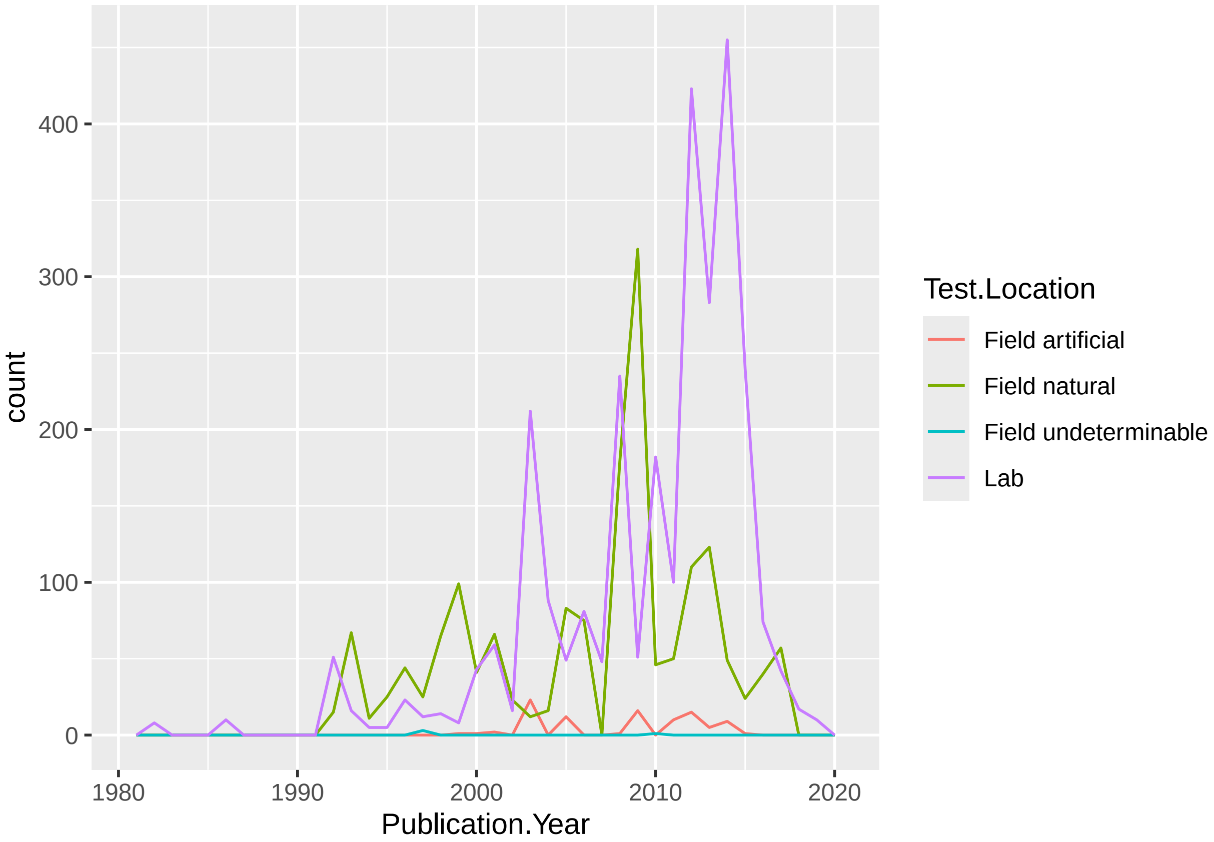
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics,aes(x=Publication.Year))+ #Data set to Neonics. X set to publication year
  geom_freqpoly(binwidth=1) #binwidth set to one to separate by year
```



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

```
ggplot(Neonics,aes(x=Publication.Year,color=Test.Location))+ #Separated out test.location by color
  geom_freqpoly(binwidth=1)
```



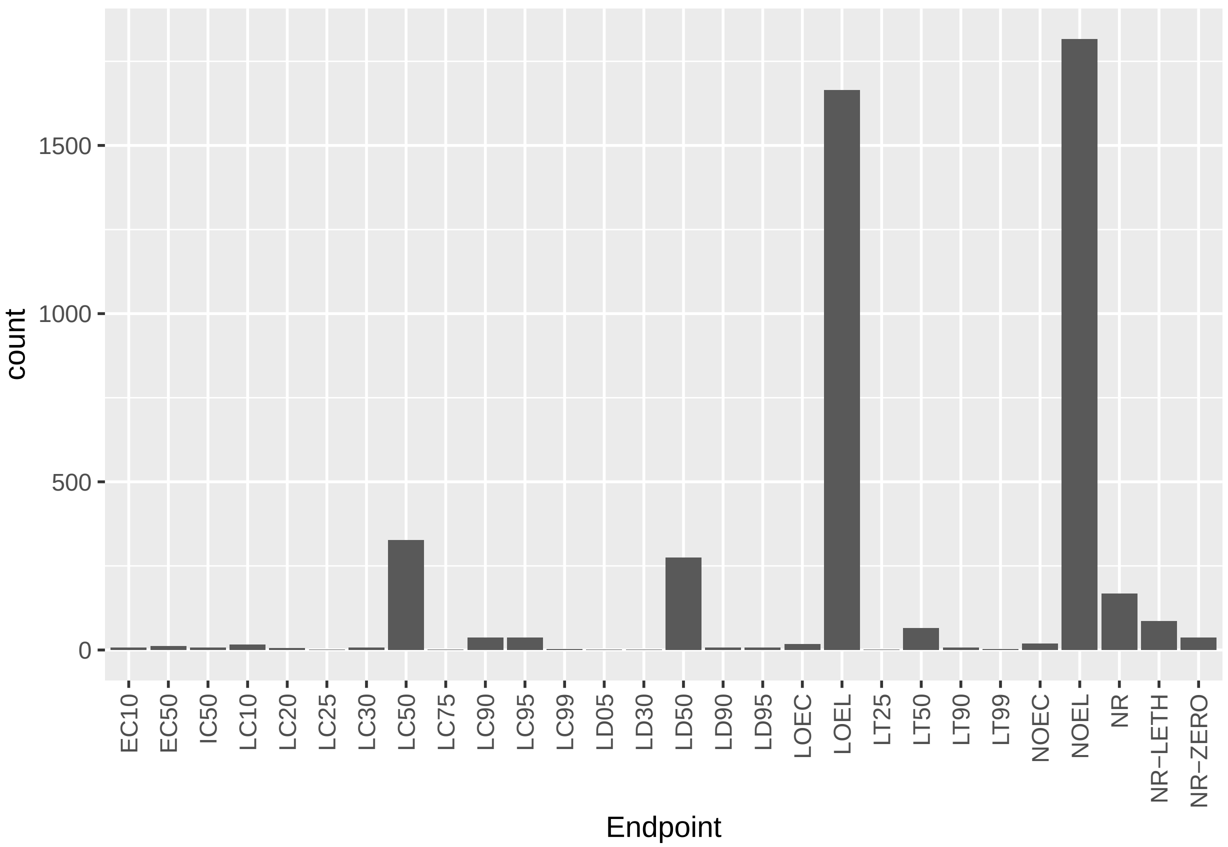
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The two most common test locations consistently over time appear to be "Lab" and "Field natural." "Field artificial" briefly appears as a test location from roughly 2000-2015 and the "Field undeterminable" location is rarely used. The "Field natural" test location peaked around 2009 and has since become less common, while the "Lab" test location increased in frequency steadily from 2000 onwards with a notable peak from 2012 to 2014.

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))+ #Data set to Neonics. X set to Endpoint
  geom_bar()+ #Bar plot
  theme(axis.text.x=element_text(angle=90,vjust=0.5,hjust=1)) #Rotate x-axis labels for visual clarity
```



Answer: The two most common end points are LOEL and NOEL. LOEL is defined as the Lowest Observed Effect Level, the lowest concentration that caused a statistically significant effect. NOEL is defined as the No Observed Effect Level, the highest concentration tested that showed no 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) #Check class of collectDate. Is factor
```

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

```
Litter$collectDate<-as.Date(Litter$collectDate) #Convert factor variable to date
unique(Litter$collectDate) #Unique date values of August 2nd and August 30th
```

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

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`?

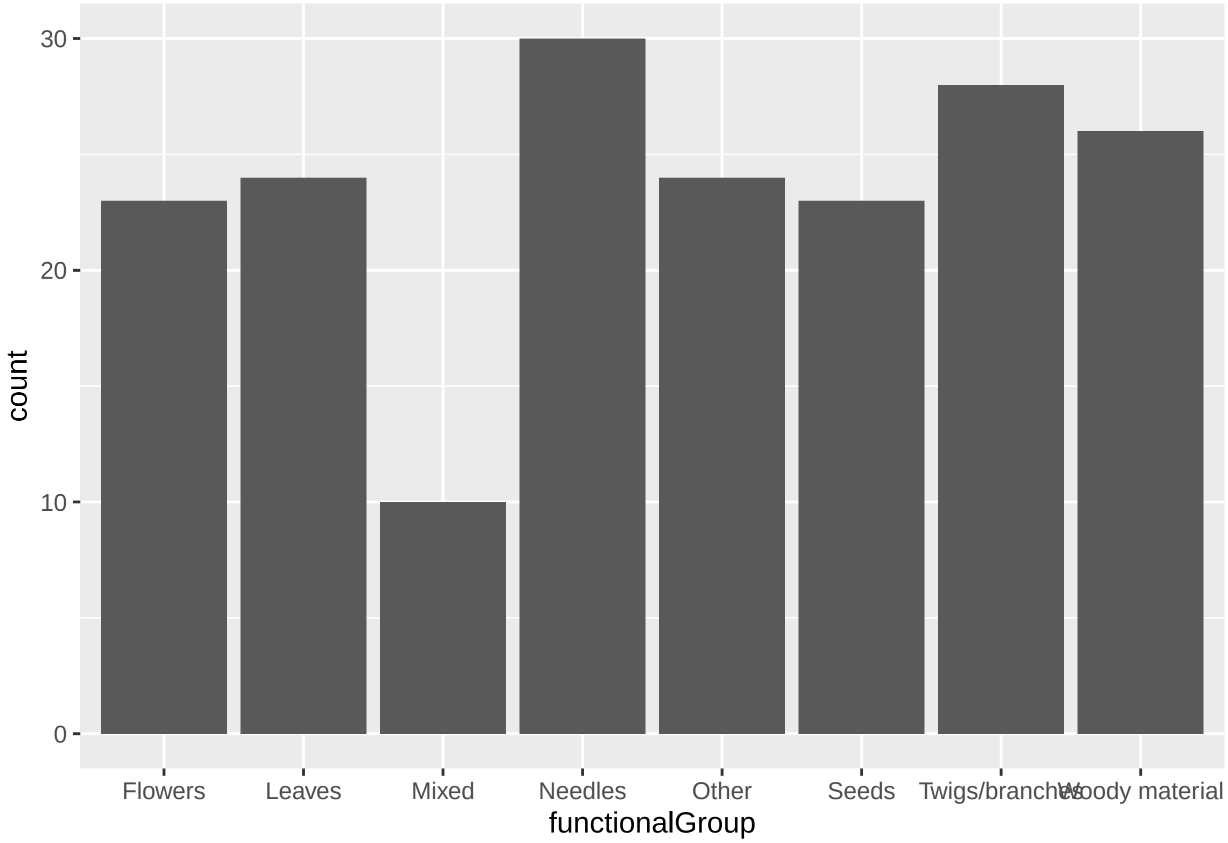
```
unique(Litter$plotID) #Twelve unique plot values found
```

```
##  [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
```

Answer: The summary function would display all levels of the variable and the count associated with the level, while unique only displays a list of the unique variables contained in the column.

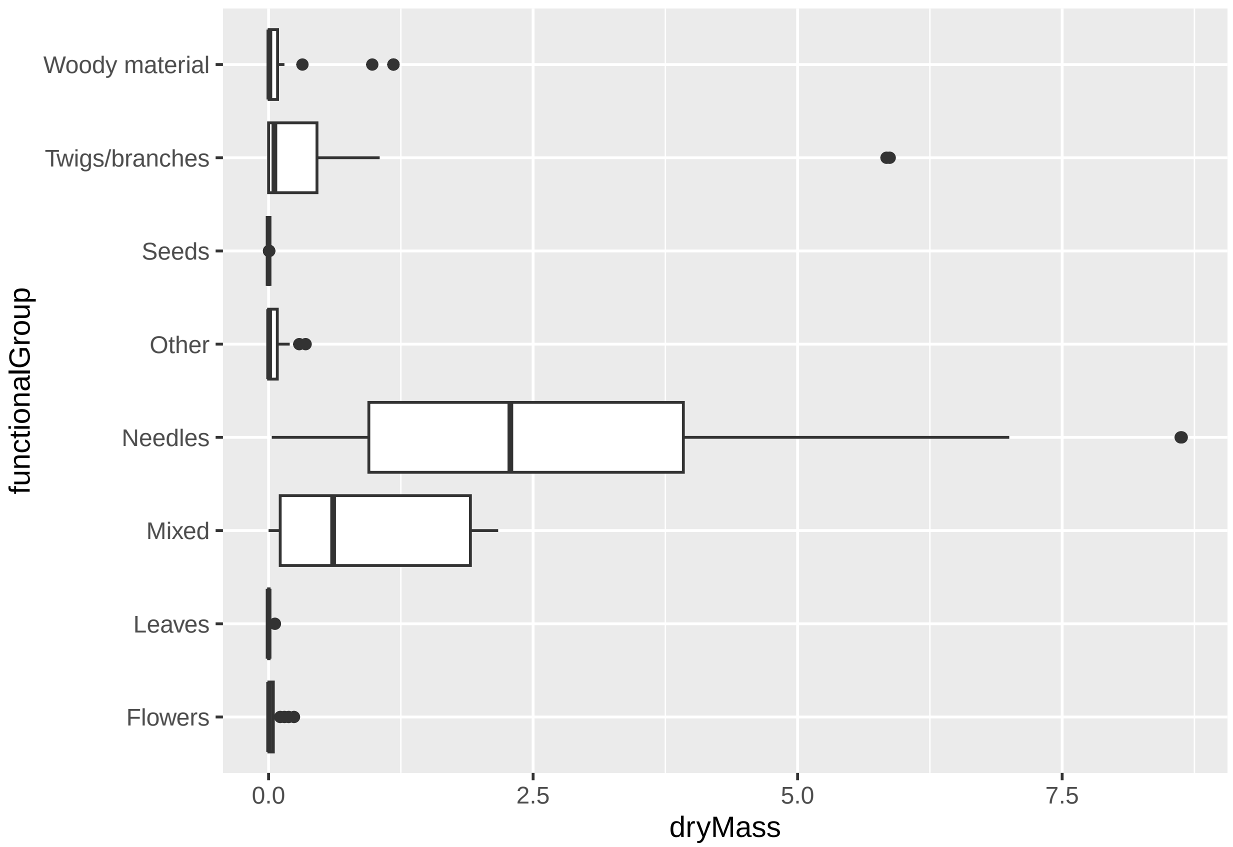
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter,aes(x=functionalGroup))+ #Data set to Litter. X set to functional group
  geom_bar() #Bar graph
```

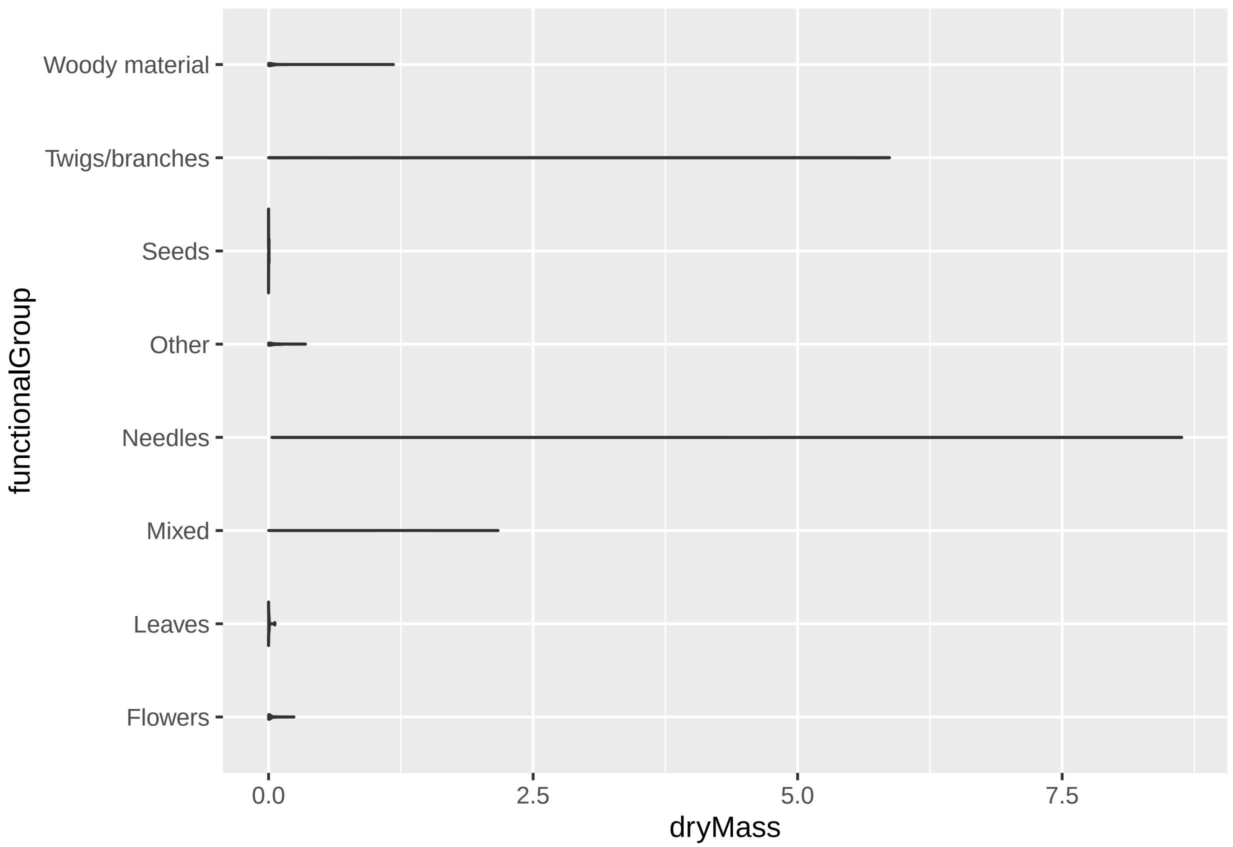


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(Litter,aes(x=dryMass,y=functionalGroup))+ #Data set to Litter. X set to dryMass. Y set to functi
  geom_boxplot()
```



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



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

> Answer: Due to the low number of total observations (188) over the spread of the data, the density estimation and distribution displayed in the violin plot is very narrow and not visually effective in transmitting information. The box plot clearly displays the median, quartiles, and outliers with ease of interpretation.

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

> Answer: The litter that tends to have the highest biomass at these sites is Needles, Mixed, then Twigs/branches.