Assignment 3: Data Exploration

Becca Cox

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)
```

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

```
getwd()
```

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

```
here()
```

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

```
setwd(here())

dir()
```

```
## [1] "Assignments"        "Data"               "EDE_Fall2024.Rproj"
## [4] "Lessons"            "README.md"
```

```
neonics <- read.csv(
  file = here("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"),
  stringsAsFactors = TRUE)

litter <- read.csv(
  file = 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 have shown neonicotinoids to be toxic to some insect species, especially bees, which are critical pollinators (Xerces Society, 2024). While neonicotinoids may protect crops from pests, they are harming insects that are necessary for the crops' reproduction and fruit development.

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 that fall to the ground in forests are an important part of the nutrient cycle as they decompose to become soil that then provides nutrients for the trees and vegetation. The decomposed matter also contributes to carbon storage (USDA, 2000).

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 woody debris are collected in traps, the contents of which are then sorted by type and measured. 2. Researchers try to always sample in the same locations. 3. Ground traps, which catch material with greater lengths, are sampled just once a year, while elevated traps, which catch shorter material, have more variation: deciduous forest sites are sampled once every two weeks, and evergreen forest sites are sampled once every one to two months.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the 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? [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
```

```
sort(summary(neonics$Effect), decreasing = TRUE)
```

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

   Answer: By identifying the most common effects, we can look more specifically at how the insecticides are impacting the insects. It can indicate that there is substantial evidence of a problem that then can be presented to stakeholders. It can also reveal effects that may need further study to better understand.

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, the Parasitic Wasp, the Buff Tailed Bumblebee, the Carniolan Honey Bee, the Bumble Bee, and the Italian Honeybee. All of these are pollinators. They are of interest because they are important to the pollination and subsequent fruit production of plants. They are economically important to farmers as the success of their harvest depends on the pollination of their crops.

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"
```

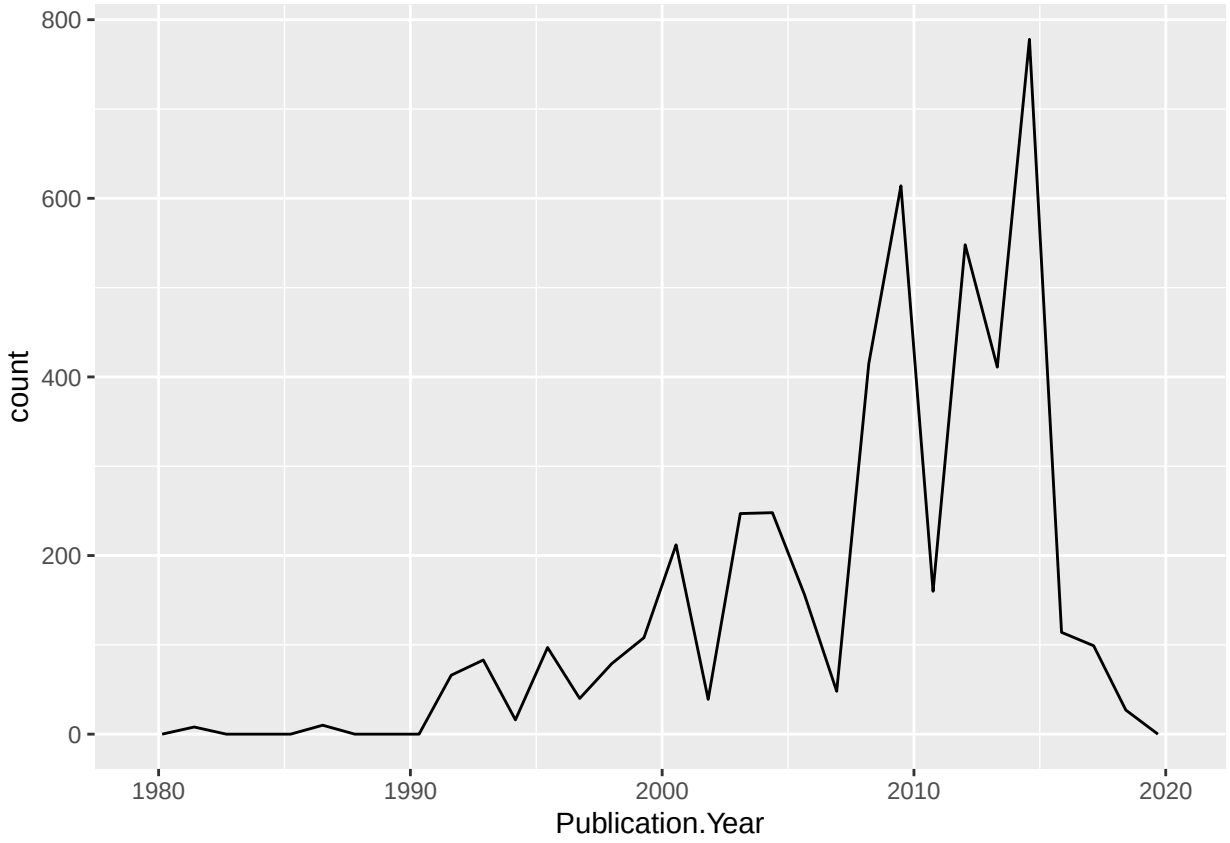
Answer: The class "Conc.1..Author" is in the factor class, not numeric, because all of the concentrations are measured in different units. You can't complete computations with the different values because they don't share units.

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))
```

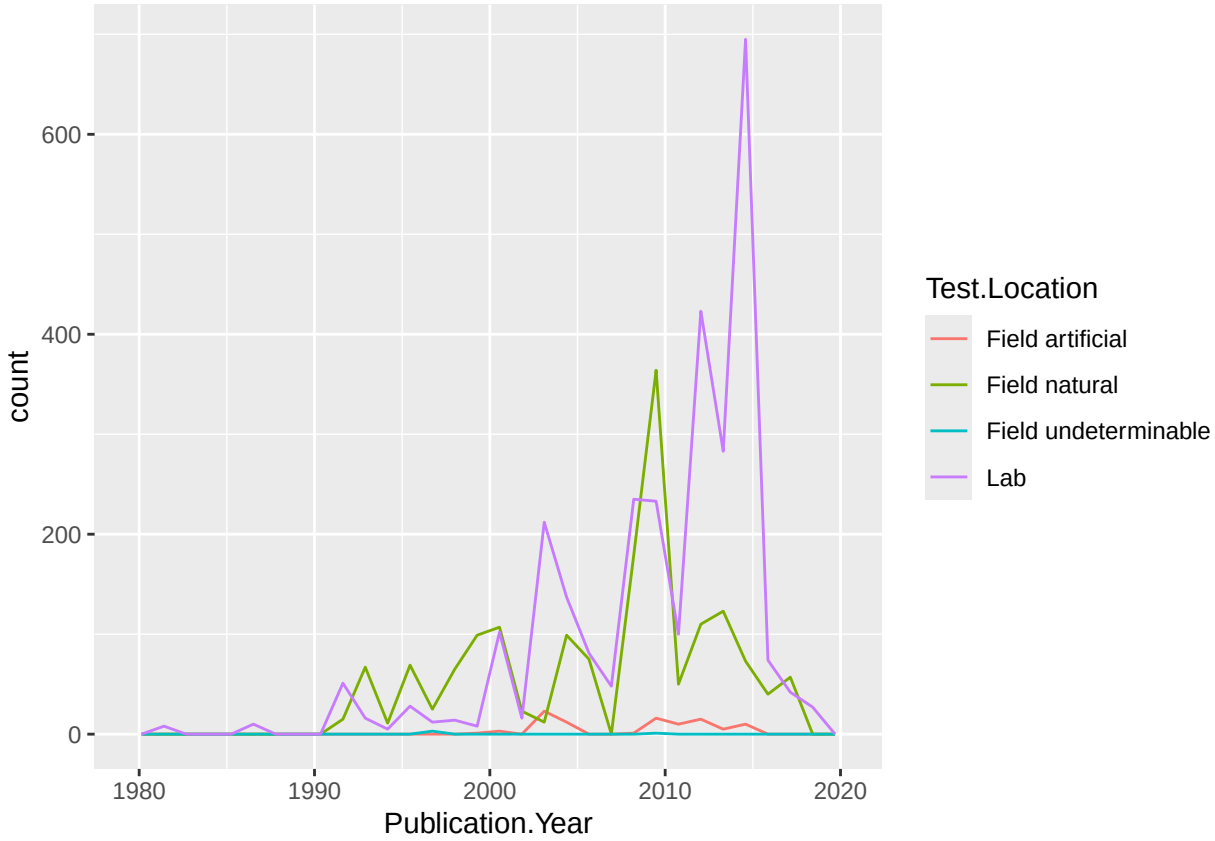
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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))
```

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




```
summary(neonics$Test.Location)
```

```
##     Field artificial        Field natural Field undeterminable
##                   96                 1663                    4
##                  Lab
##                 2860
```

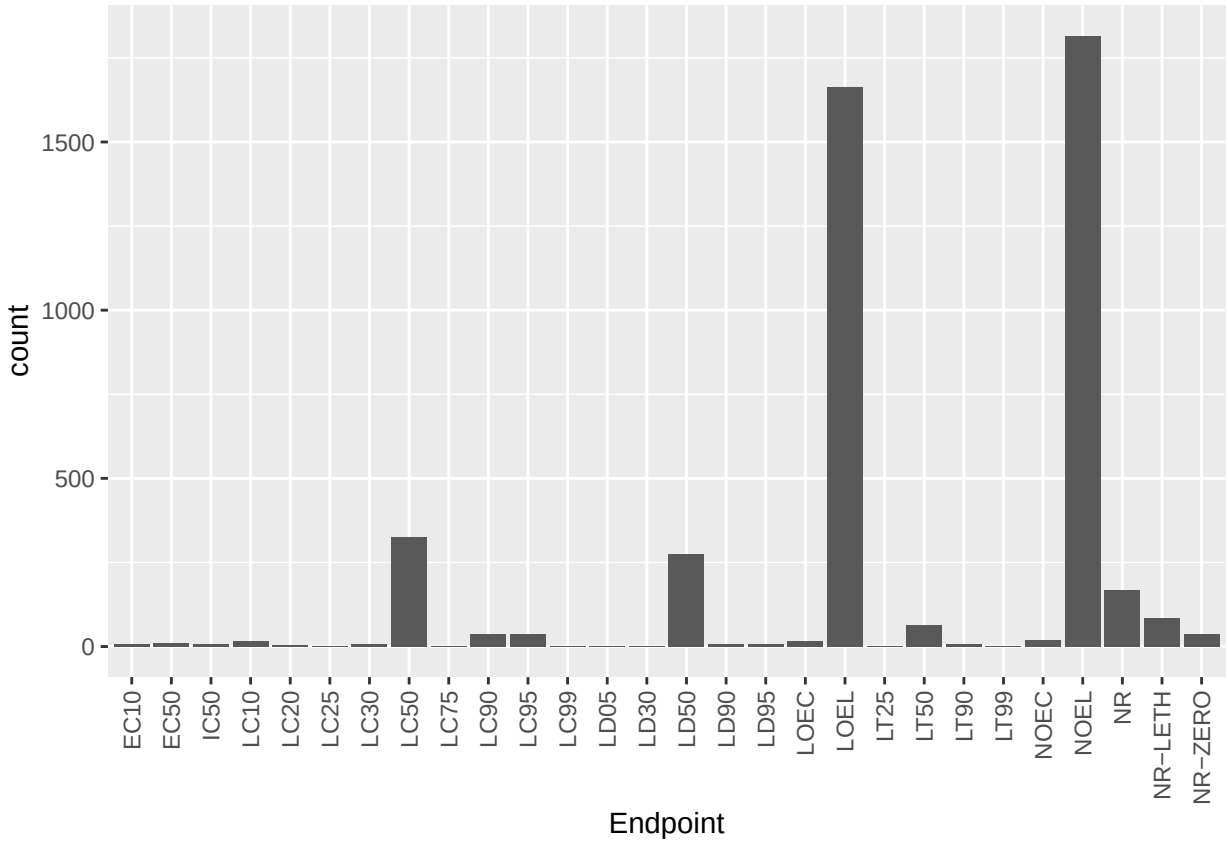
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: Most of the tests (n = 2860) took place in a lab. The second most common test location is in the Field (natural) (n=1663). There were not many studies taking place until the 1990s, when the number of tests in the lab and in the Field (natural) were comparable until the early 2000s. Around 2002, there was a spike in Lab studies. At this time, there is also a slight increase in the number of Field (artificial) studies. The total number of studies being conducted dropped quickly around 2007. The number of Lab and Field (natural) studies increased again until 2009, then dropped again until late 2010. The number of Field (artificial) studies also had a slight increase beginning around the same time, and it was steady, yet still low, until 2016. After late 2010, the number of Field (natural) studies remained lower and continued decreasing overall. The number of Lab studies increased overall, peaking at almost 700 studies in late 2014, then declined quickly until 2016. Since 2016, both Lab and Field (natural) studies have been decreasing fairly gradually to almost no studies being produced in a year.

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



Answer: The two most common endpoints are NOEL and LOEL. NOEL means that the study found that the highest concentration of the insecticide did not cause responses that were significantly different from the responses of the control group. LOEL, on the other hand, means that the lowest concentration caused effects that were signficiantly different than those experienced by the control group.

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"
```

```
litter$collectDate <- ymd(litter$collectDate)

class(litter$collectDate)
```

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

```
unique(litter$collectDate)
```

```
## [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)
```

```
##  [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
```

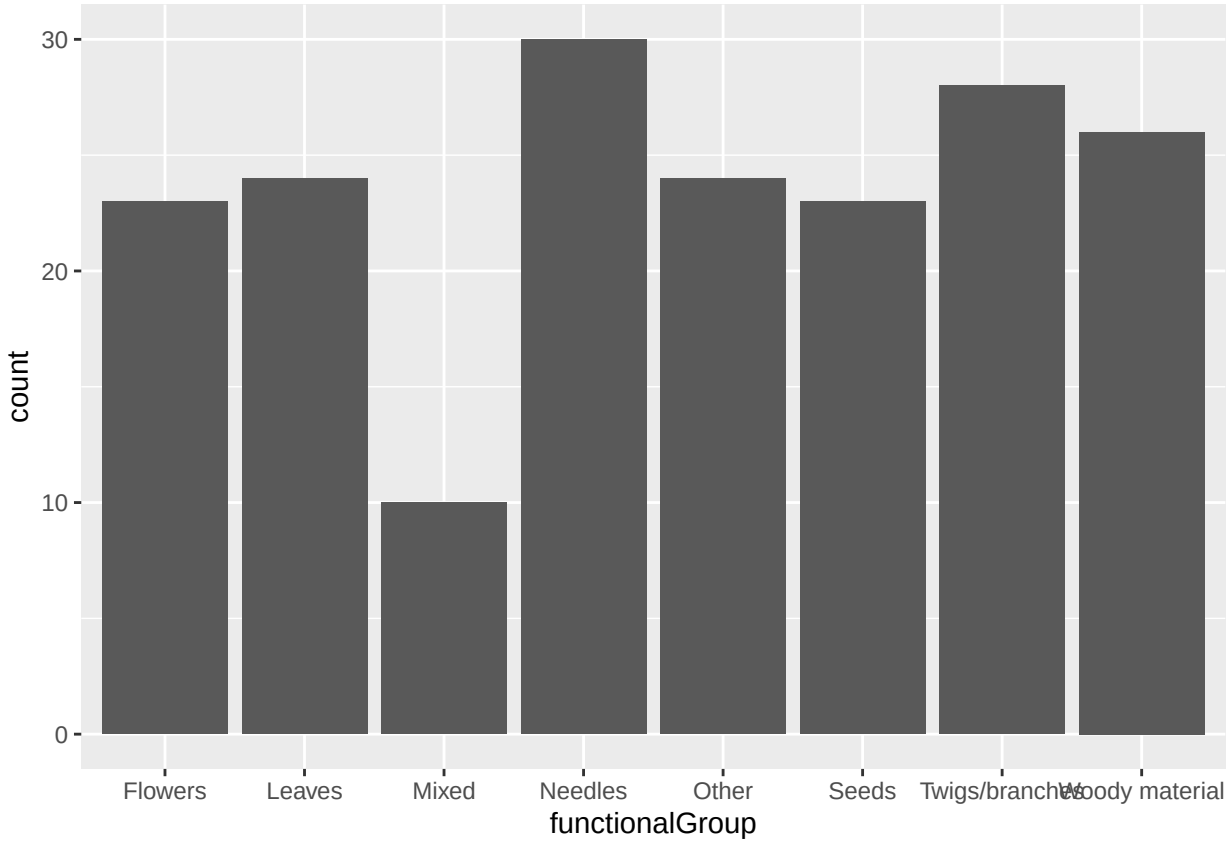
```
summary(litter$plotID)
```

```
## 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: The 'unique' function lists the different plots (levels) and tells you how many there are. The 'summary' function tells you how many counts of each of the different plots (levels) there are in the dataset.

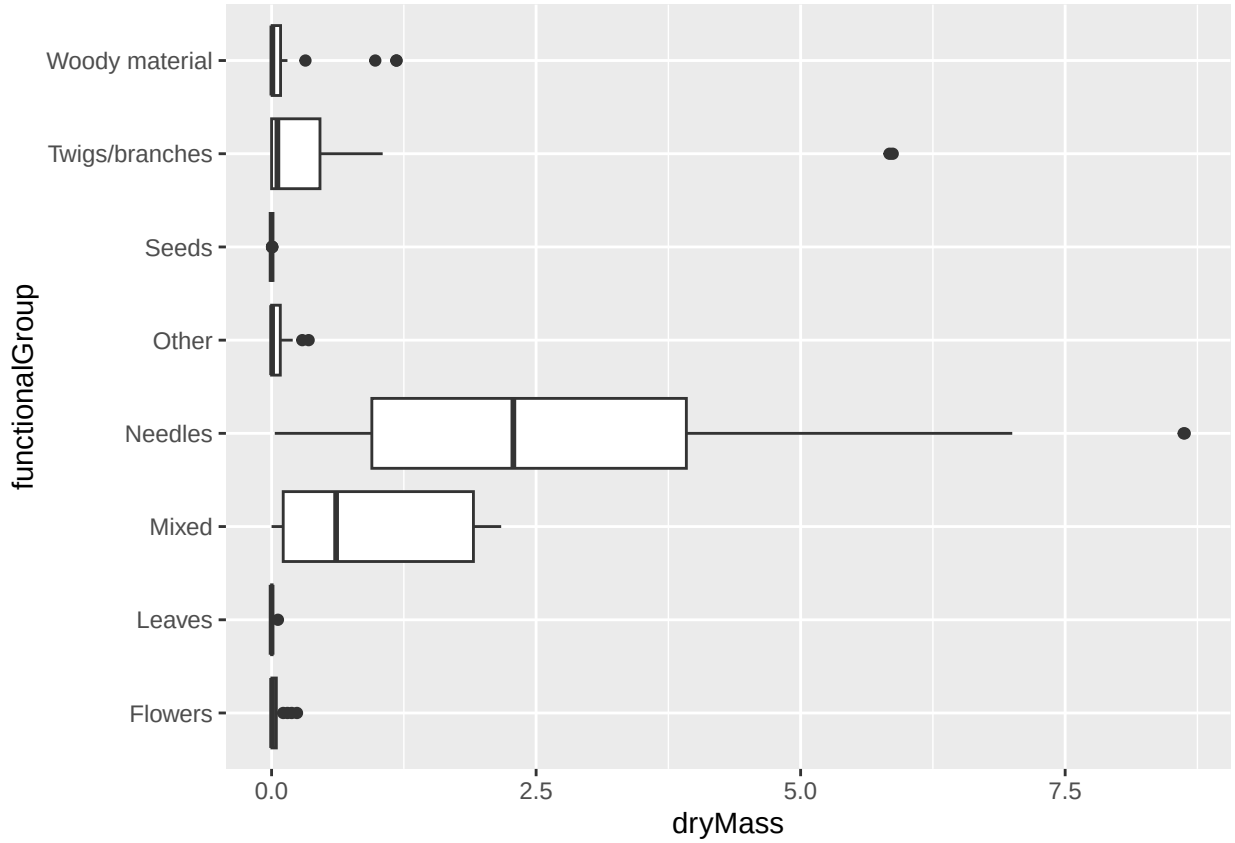
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

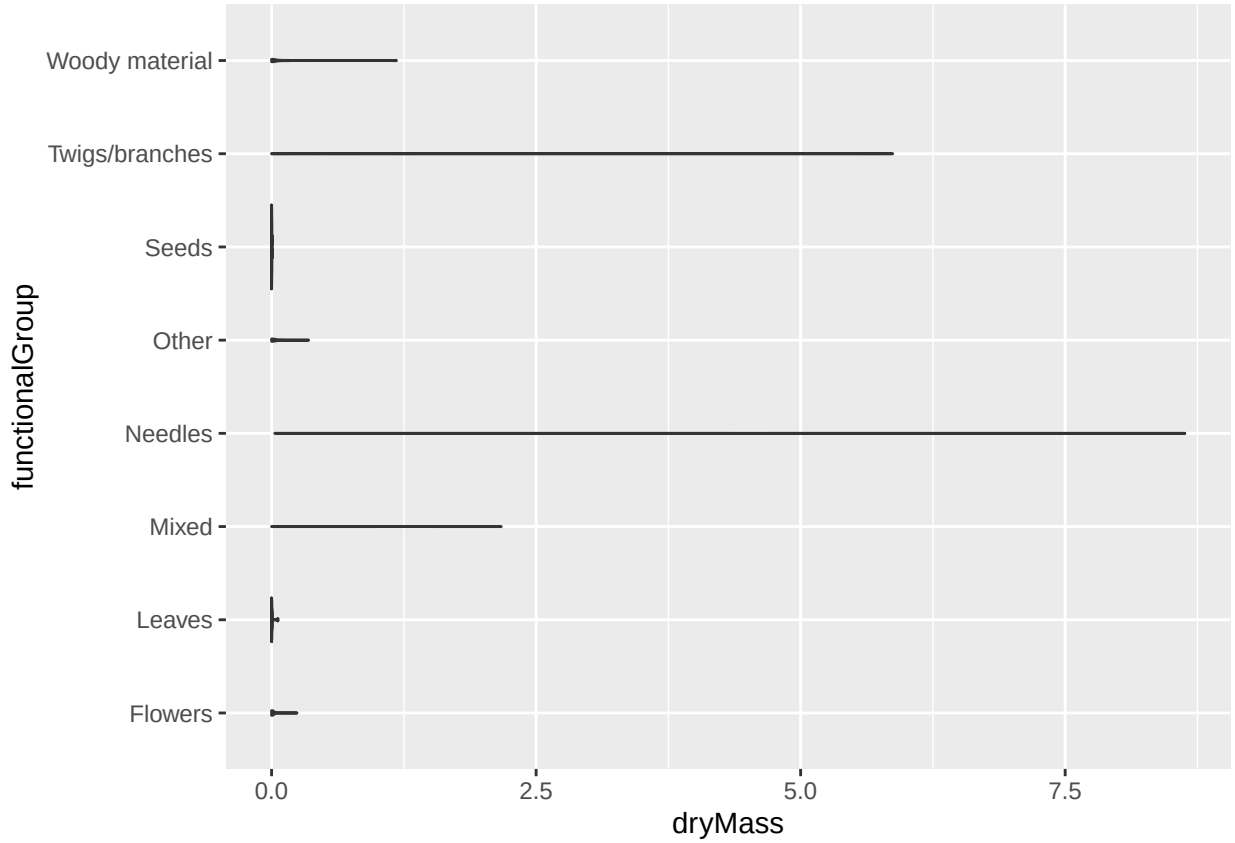


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




```
ggplot(litter) +
  geom_violin(aes(x = dryMass, y = functionalGroup),
              draw_quantiles = c(0.25, 0.5, 0.75))
```



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

> Answer: The violin plot with this data is hard to read: violin plots show the shape of the distribution of data, and these data appear to be either very evenly spread out across dryMass values, or highly concentrated at very low dryMass values. There is not a clear shape for any of the functionalGroups. The boxplot highlights the middle 50% of the data as well as the range, and includes outliers. This gives a better sense of where the data lie across the dryMass values. It is more effective because it is clearer than the violin plot, distinguishing characteristics of the data that are not visible in a violin plot.

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

> Answer: Needles tend to have the highest biomass, with the highest average (~2.25) and greatest range of values. The second highest type, mixed litter, has an average (~0.625) that is much lower, though its range overlaps with that of needles.