Assignment 3: Data Exploration

Hannah Wudke

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.

```
#1 Here, I am installing and loading the necessary packages:
install.packages("tidyverse")
install.packages("lubridate")
install.packages("here")

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

# Here, I am checking my current working directory
getwd()
```

```
## [1] "/home/guest/EDE-Class"
```

```
#1a Here, I am uploading and naming two datasets:
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: Neonicotinoids are a broad group of insecticides used in applications ranging from flea and tick treatments in pets to agricultural insect management. Although they are very effective for these applications, they are particularly toxic to bees. Because of this unintended consequence on pollinator populations, neonics have been banned in several European countries. We may be interested in the ecotoxicology of neonics to understand their costs and benefits, with respect to declining pollinator populations and agricultural benefits.

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 is a crucial part of forest ecosystem health. It provides habitats for insects and small animals, affects water and sediment transport, impacts the local carbon budget, and plays into nutrient cycling. It can also be impacted by flooding and fire events. In Colorado in particular, forest fires likely play a significant role in the presence of woody debris and litter.

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. In deciduous forests, samples are taken once every two weeks. In evergreen forests, samples are only taken once per every 1-2 months. Some sites may not be sampled for up to six months if access is limited, such as in the winter months. 2. Sampling is conducted in tower plots, which are randomly selected in primary and secondary airsheds. The size of the plots depends on vegetation types. 3. One elevated litter trap and one ground litter trap are used in every plot area, and there are 1-4 pairs of these traps per plot (400m^2).

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
view(Neonics)
class(Neonics)
```

```
## [1] "data.frame"
```

```r
colnames(Neonics)
```

```
##  [1] "CAS.Number"                        "Chemical.Name"
##  [3] "Chemical.Grade"                    "Chemical.Analysis.Method"
##  [5] "Chemical.Purity"                   "Species.Scientific.Name"
##  [7] "Species.Common.Name"               "Species.Group"
##  [9] "Organism.Lifestage"                "Organism.Age"
## [11] "Organism.Age.Units"                "Exposure.Type"
## [13] "Media.Type"                        "Test.Location"
## [15] "Number.of.Doses"                   "Conc.1.Type..Author."
## [17] "Conc.1..Author."                   "Conc.1.Units..Author."
## [19] "Effect"                            "Effect.Measurement"
## [21] "Endpoint"                          "Response.Site"
## [23] "Observed.Duration..Days."          "Observed.Duration.Units..Days."
## [25] "Author"                            "Reference.Number"
## [27] "Title"                             "Source"
## [29] "Publication.Year"                  "Summary.of.Additional.Parameters"
```

```r
dim(Neonics)
```

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

```r
length(Neonics)
```

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

```r
# Here, I broadly explored the data set. Then, I ran dim(Neonics),
# which showed 4623 entries 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...]

```r
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 that are studied are Mortality and Population. These have >1000 occurances. Also common are Behavior, Feeding behavior, and Reproduction. Mortality and population are arguably the most important markers of neonicotinoids impacting the ecosystem. If the neonics are killing the pollinators and reducing the population size, the impact on plants and biodiversity will be significant. This would be of interest to ecotoxicologists and ecologists.

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)
```

```
##                          Honey Bee                     Parasitic Wasp
##                                667                                285
##              Buff Tailed Bumblebee                Carniolan Honey Bee
##                                183                                152
##                         Bumble Bee                   Italian Honeybee
##                                140                                113
##                    Japanese Beetle                  Asian Lady Beetle
##                                 94                                 76
##                     Euonymus Scale                           Wireworm
##                                 75                                 69
##                  European Dark Bee                  Minute Pirate Bug
##                                 66                                 62
##                Asian Citrus Psyllid                      Parastic Wasp
##                                 60                                 58
##             Colorado Potato Beetle                    Parasitoid Wasp
##                                 57                                 51
##                Erythrina Gall Wasp                      Beetle Order
##                                 49                                 47
##         Snout Beetle Family, Weevil           Sevenspotted Lady Beetle
##                                 47                                 46
##                      True Bug Order              Buff-tailed Bumblebee
##                                 45                                 39
##                       Aphid Family                    Cabbage Looper
##                                 38                                 38
##                Sweetpotato Whitefly                    Braconid Wasp
##                                 37                                 33
##                      Cotton Aphid                    Predatory Mite
##                                 33                                 33
##              Ladybird Beetle Family                       Parasitoid
##                                 30                                 30
##                      Scarab Beetle                    Spring Tiphia
##                                 29                                 29
##                       Thrip Order                Ground Beetle Family
##                                 29                                 27
##                  Rove Beetle Family                    Tobacco Aphid
##                                 27                                 27
##                      Chalcid Wasp              Convergent Lady Beetle
##                                 25                                 25
##                     Stingless Bee                  Spider/Mite Class
```

```
##                                25                                24
##                 Tobacco Flea Beetle                 Citrus Leafminer
##                                24                                23
##                     Ladybird Beetle                        Mason Bee
##                                23                                22
##                           Mosquito                    Argentine Ant
##                                22                                21
##                              Beetle          Flatheaded Appletree Borer
##                                21                                20
##                Horned Oak Gall Wasp                 Leaf Beetle Family
##                                20                                20
##                   Potato Leafhopper          Tooth-necked Fungus Beetle
##                                20                                20
##                       Codling Moth           Black-spotted Lady Beetle
##                                19                                18
##                       Calico Scale                 Fairyfly Parasitoid
##                                18                                18
##                        Lady Beetle              Minute Parasitic Wasps
##                                18                                18
##                          Mirid Bug                  Mulberry Pyralid
##                                18                                18
##                           Silkworm                  Vedalia Beetle
##                                18                                18
##               Araneoid Spider Order                       Bee Order
##                                17                                17
##                     Egg Parasitoid                     Insect Class
##                                17                                17
##            Moth And Butterfly Order        Oystershell Scale Parasitoid
##                                17                                17
## Hemlock Woolly Adelgid Lady Beetle               Hemlock Wooly Adelgid
##                                16                                16
##                              Mite                       Onion Thrip
##                                16                                16
##                Western Flower Thrips                     Corn Earworm
##                                15                                14
##                   Green Peach Aphid                        House Fly
##                                14                                14
##                           Ox Beetle                 Red Scale Parasite
##                                14                                14
##                  Spined Soldier Bug               Armoured Scale Family
##                                14                                13
##                    Diamondback Moth                     Eulophid Wasp
##                                13                                13
##                   Monarch Butterfly                     Predatory Bug
##                                13                                13
##                Yellow Fever Mosquito                Braconid Parasitoid
##                                13                                12
##                       Common Thrip        Eastern Subterranean Termite
##                                12                                12
##                             Jassid                       Mite Order
##                                12                                12
##                          Pea Aphid                   Pond Wolf Spider
##                                12                                12
##            Spotless Ladybird Beetle              Glasshouse Potato Wasp
```

```
##                              11                                  10
##                        Lacewing           Southern House Mosquito
##                              10                                  10
##          Two Spotted Lady Beetle                         Ant Family
##                              10                                   9
##                      Apple Maggot                           (Other)
##                               9                                 670
```

Answer: The six most common species (in common names) are Honey Bees (n=667), Parasitic Wasps (n=285), Buff Tailed Bumblebee (n=183)*, Carniolan Honey Bees (n=152), Bumble Bees (n=140) and Italian Honeybees (n=113). The commonality between all of these species is that they are bees (n=5) and wasps (n=1). The other studied insects are more specific types of bees and wasps, or other types of insects like beetles, moths, worms, and ants. Bees and wasps are of significant interest over other insects because they often play a more crucial role in the ecosystem. Bees and Wasps are pollinators, and for this reason, have a major role in plant/crop production.

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)
class(Neonics$Conc.1..Author.)
```

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

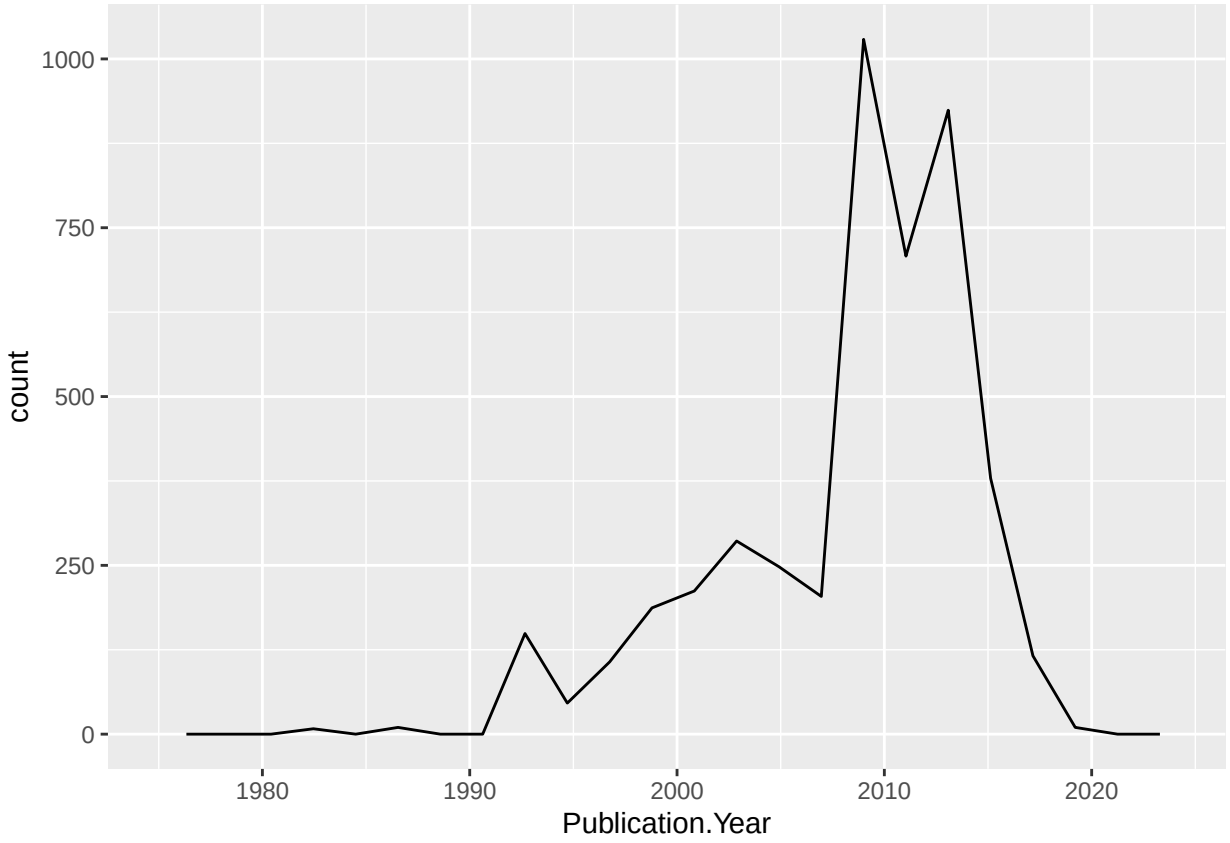
Answer: The class of Conc.1..Author. is a factor. This means that each variable has a fixed and known set of values. Generally, factors are categorical. The three Conc. factors are referring to the type of ingredient (active, formulation), concentration, and units, respectively. Some of the concentration values contain "/" after the number. The addition of characters such as "/" means the data cannot be listed as numeric, it must be a factor.

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),
                          bins = 25) + scale_x_continuous(limits = c(1975, 2024))
```

```
## Warning: Removed 3 rows containing missing values or values outside the scale range
## (`geom_path()`).
```



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),
                            #    bins = 25) + geom_freqpoly(aes(x = Test.Location),
                                                          #  bins = 25, color = "red")
# I am having significant issues with these two graphs because of the different
# classes of the data. Will discuss in office hours!
```

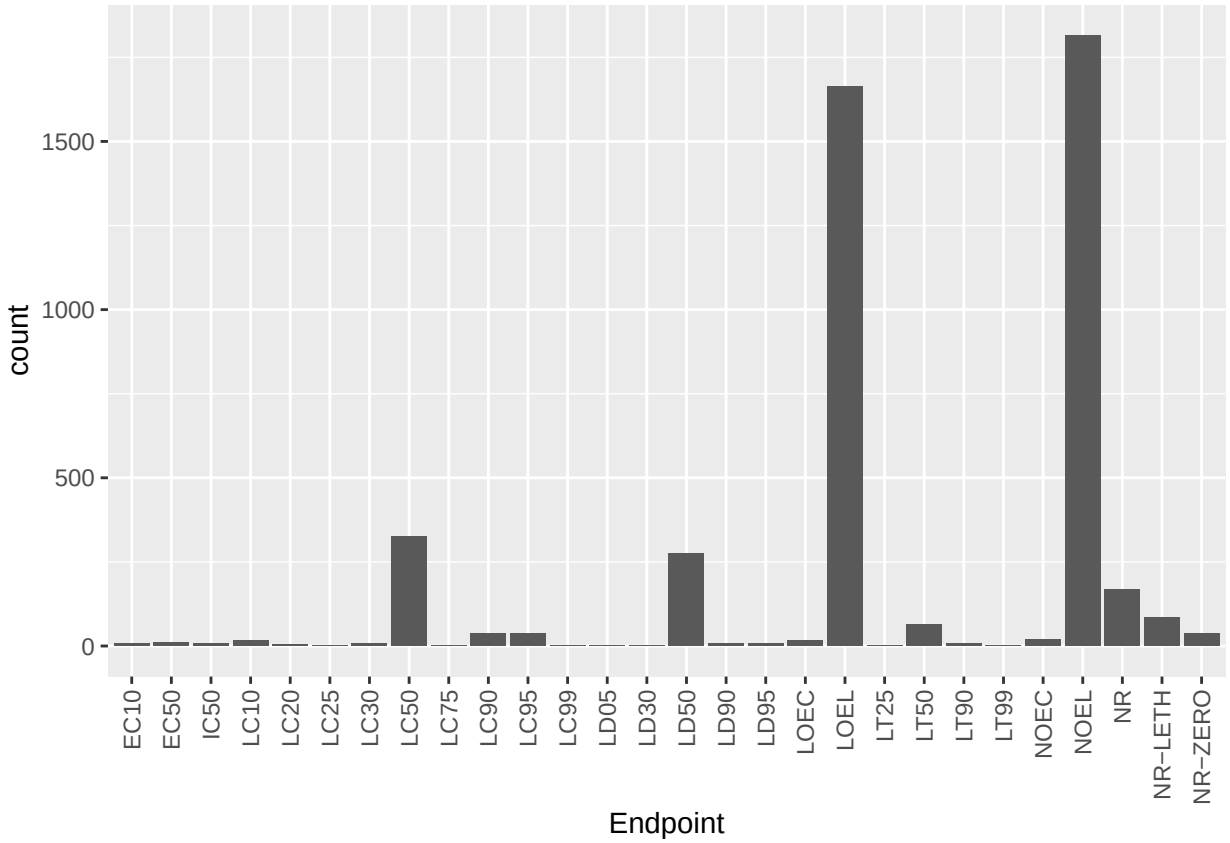
Interpret this graph. What are the most common test locations, and do they differ over time?

Answer:

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. LOEL means lowest observable effect level, which is the lowest possible dose that produces effects different from the control group. This is in terrestial systmes. NOEL means no observable effect level, which means the lowest possible dose that produces no effect different from 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"
```

```
# view(Litter)
# It is not currently a date, it is a factor. Let's change that!
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
class(Litter$collectDate)
```

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

```r
unique(Litter$collectDate)
```

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

```r
# This showed that in August 2018, litter was collected on the 2nd and 30th
```

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`?

```r
unique(Litter$collectDate)
```

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

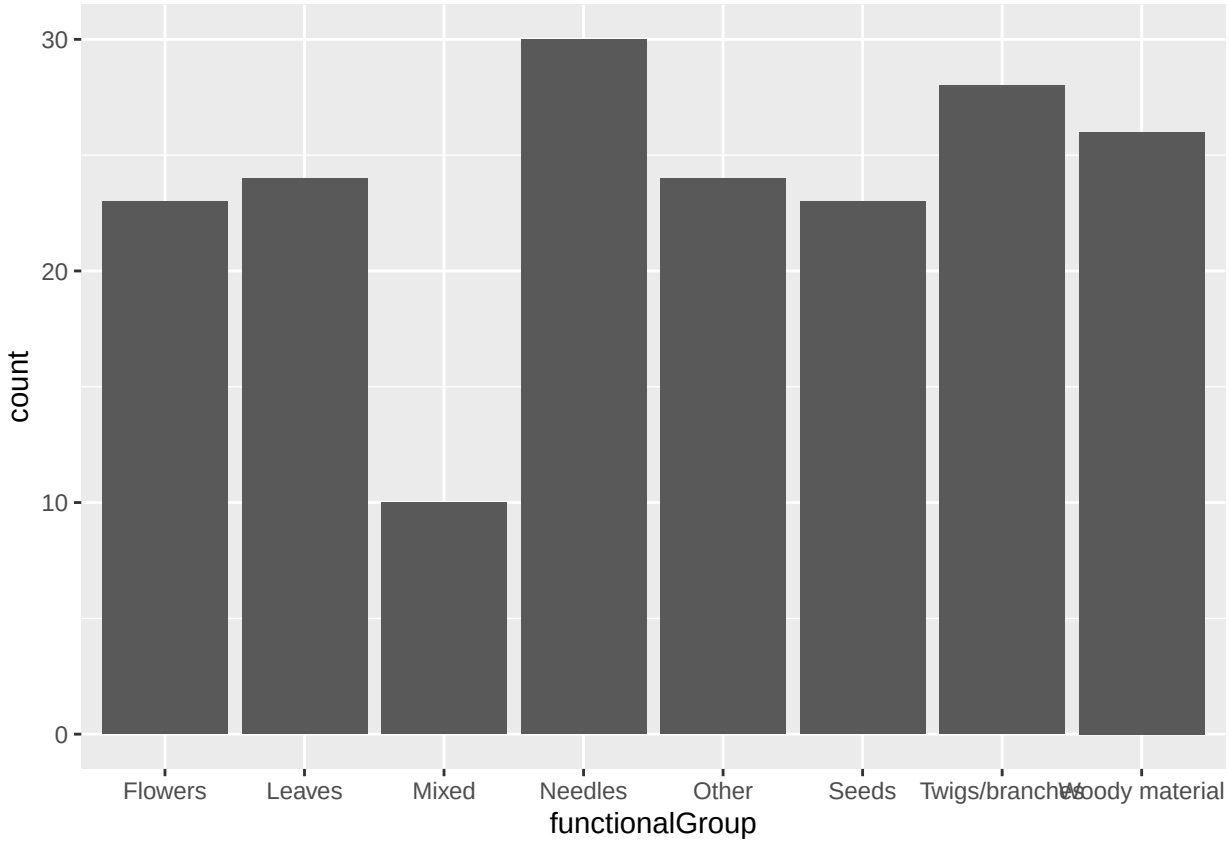
```r
summary(Litter$collectDate)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2018-08-02" "2018-08-02" "2018-08-30" "2018-08-16" "2018-08-30" "2018-08-30"
```

Answer: The unique function shows each unique variable that is reported. In this case, there are two, 2018-08-02 ans 2018-08-30. These are the only two values, although they occur many times each. Summary shows the minimum, median, mean, maximum, and 1st and 3rd quartiles. Although most of these parameters yield 2018-08-30 and 2018-08-02, the mean is calculated, and yields 2018-08-16. This mean does not actually occur in the dataset, which could cause issues in data interpretation if summary and unique were not used together.

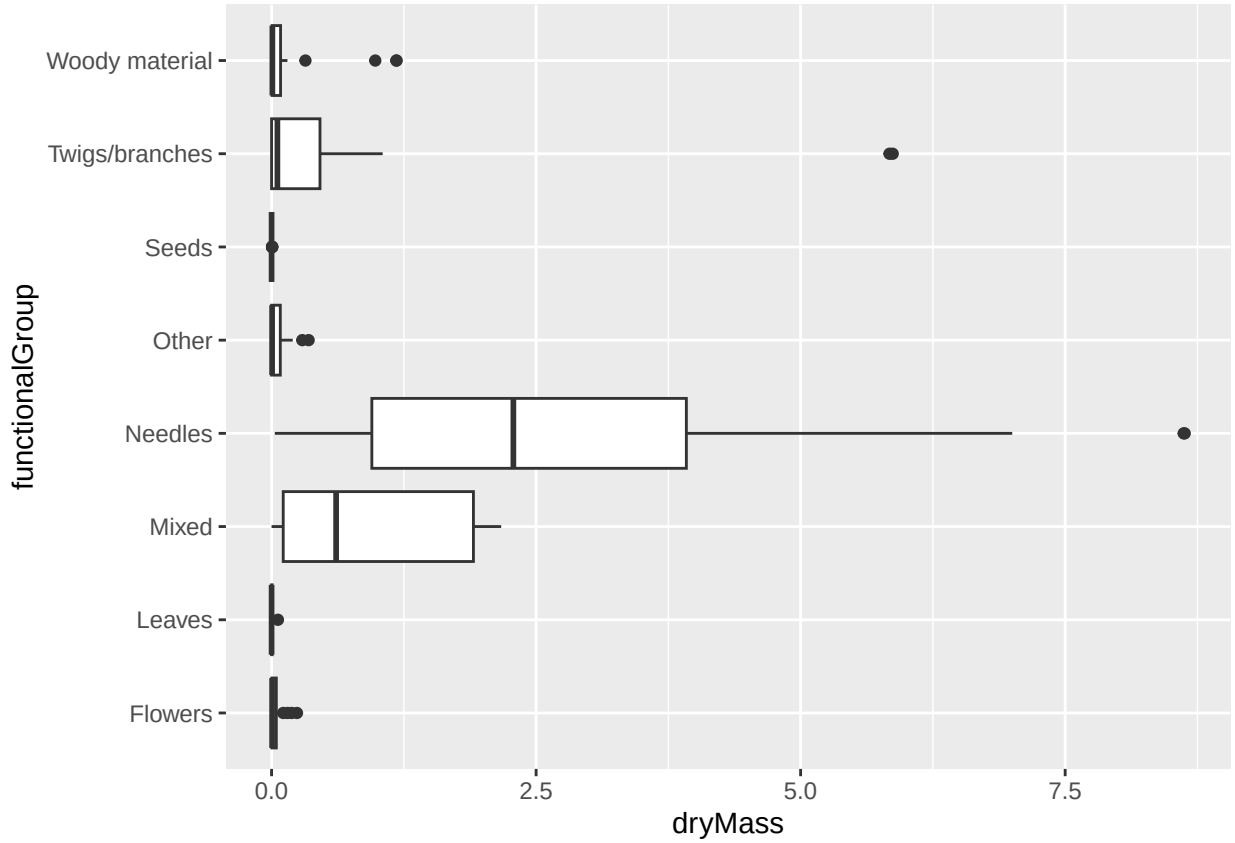
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

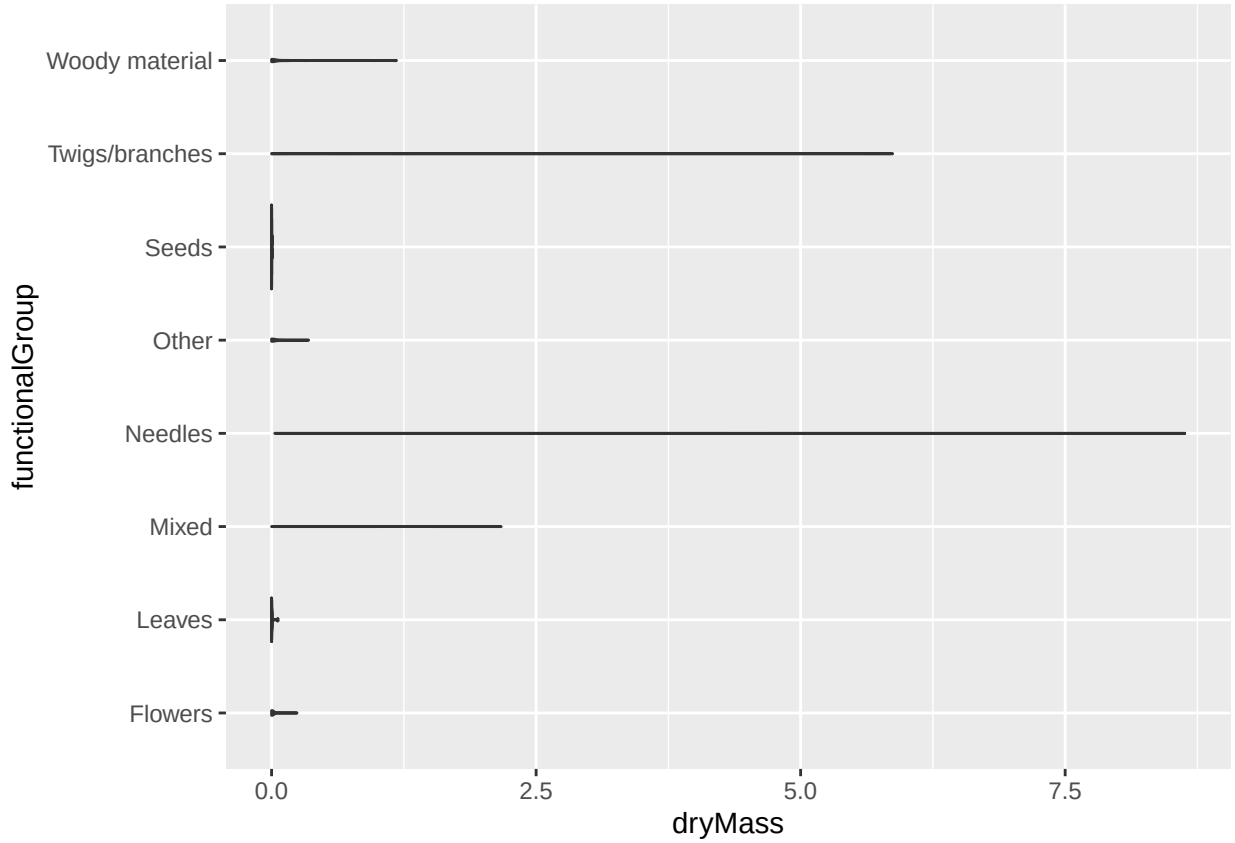


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




```
ggplot(Litter) + geom_violin(aes(x = dryMass, y = functionalGroup),
                             draw_quantiles = c(0.25, 0.50, 0.75))
```



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

> Answer: Boxplots are much more effective at showing data distribution, rather than just maximum/range. For instance, the median can be a fast, effective way to determine which groups occur the most relative to others. This magnitude is something that is lost with the violin plot.

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

> Answer: Needles. Although there are some isolated examples of twigs/branches and mixed groups that have comparable dry mass, the median of needles far exceeds all of the medians of the other data sets.