Assignment 3: Data Exploration

Emma Childs

Spring 2023

#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.

```
library(tidyverse)
library(lubridate)

ECOTOX_Neonicotinoids_Insects_raw <- read_csv("Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv")

Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",stringsAsFactors = TRUE)

NEON_NIWO_Litter_massdata_2018_08_raw <- read_csv("Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv")


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

#Imported both datasets and then renamed as Neonics and Litter, using the string as factors command.
```

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: Oh, I'm so interested in this because of my background in agriculture! There are lots of insects that are incredibly beneficial for farming and other agricultural projects - some of them are critical and essential for pollination of plants and other ecosystem benefits. If neonicotinoids are harmful for these insects and impact their ability to function, reproduce, fly, etc., it could be incredibly detrimental for many plants and the whole ecosystem (humans included)

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: Studying leaf litter and woody debris would yield data into how the neonicotinoids might bioaccumulate in plant matter, which would have implications for forest insects who eat or live in the plant matter. If different species are either living in the trees, the forest floor, or underground then their life cycles could be affected by these potentially lingering insecticides.

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. For this data collection, the scientists designated litter as material that is dropped from the forest canopy and had a diameter <2cm and a length <50 cm. The material was collected in elevated 0.5m2 PVC traps. Fine wood debris was defined as material that dropped from the forest canopy and had a diameter <2cm and a length >50 cm. This material was collected in ground traps as longer material was not reliably collected by the elevated traps. 2. According to the user guide, trap placement within plots were either targeted or randomized, depending on the vegetation. I found this interesting because it is not typical spatial sampling, but I imagine it would allow for greater examination of any particular areas that scientists thought might be especially impacted by neonicotinoids, or particular species of concern. In reading further, I see that the targeted plots are chosen based on lower percent vegetation cover, which makes sense. 3. Time of year is incredible important for this sampling design, as both deciduous and evergreen areas are being evaluated. At sites with deciduous vegetation or limited access during winter months, the guide says litter sampling of elevated traps may be discontinued for up to 6 months during the dormant season.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

```
#4623x30 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?

```
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:Population and Mortality are by far the two most commonly studied effects. This makes sense because they would provide measurable insight into how the neonicotinoid is impacting the insects. If their population is decreasing or their mortality rates are affected, that could point to some degree of connection with the insecticide use. Are they alive? Are their numbers going up or down? Clear questions that can be studied.

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…]

```
sort(summary(Neonics$Species.Common.Name))
```

```
##                         Ant Family                       Apple Maggot
##                                  9                                  9
##              Glasshouse Potato Wasp                           Lacewing
##                                 10                                 10
##             Southern House Mosquito            Two Spotted Lady Beetle
##                                 10                                 10
##            Spotless Ladybird Beetle                Braconid Parasitoid
##                                 11                                 12
##                       Common Thrip      Eastern Subterranean Termite
##                                 12                                 12
##                             Jassid                         Mite Order
##                                 12                                 12
##                          Pea Aphid                   Pond Wolf Spider
##                                 12                                 12
##               Armoured Scale Family                    Diamondback Moth
##                                 13                                 13
##                      Eulophid Wasp                   Monarch Butterfly
##                                 13                                 13
##                      Predatory Bug                Yellow Fever Mosquito
##                                 13                                 13
```

```
##                 Corn Earworm                Green Peach Aphid
##                           14                               14
##                    House Fly                        Ox Beetle
##                           14                               14
##             Red Scale Parasite               Spined Soldier Bug
##                           14                               14
##         Western Flower Thrips Hemlock Woolly Adelgid Lady Beetle
##                           15                               16
##         Hemlock Wooly Adelgid                             Mite
##                           16                               16
##                  Onion Thrip             Araneoid Spider Order
##                           16                               17
##                    Bee Order                   Egg Parasitoid
##                           17                               17
##                 Insect Class          Moth And Butterfly Order
##                           17                               17
## Oystershell Scale Parasitoid          Black-spotted Lady Beetle
##                           17                               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
##                 Codling Moth       Flatheaded Appletree Borer
##                           19                               20
##          Horned Oak Gall Wasp                Leaf Beetle Family
##                           20                               20
##            Potato Leafhopper       Tooth-necked Fungus Beetle
##                           20                               20
##                Argentine Ant                           Beetle
##                           21                               21
##                    Mason Bee                         Mosquito
##                           22                               22
##             Citrus Leafminer                   Ladybird Beetle
##                           23                               23
##            Spider/Mite Class               Tobacco Flea Beetle
##                           24                               24
##                 Chalcid Wasp            Convergent Lady Beetle
##                           25                               25
##                Stingless Bee              Ground Beetle Family
##                           25                               27
##             Rove Beetle Family                   Tobacco Aphid
##                           27                               27
##                Scarab Beetle                    Spring Tiphia
##                           29                               29
##                  Thrip Order            Ladybird Beetle Family
##                           29                               30
##                   Parasitoid                    Braconid Wasp
##                           30                               33
##                 Cotton Aphid                    Predatory Mite
##                           33                               33
```

```
##                 Sweetpotato Whitefly                                Aphid Family
##                                   37                                          38
##                       Cabbage Looper                       Buff-tailed Bumblebee
##                                   38                                          39
##                       True Bug Order                    Sevenspotted Lady Beetle
##                                   45                                          46
##                         Beetle Order                Snout Beetle Family, Weevil
##                                   47                                          47
##                   Erythrina Gall Wasp                             Parasitoid Wasp
##                                   49                                          51
##               Colorado Potato Beetle                               Parastic Wasp
##                                   57                                          58
##                 Asian Citrus Psyllid                           Minute Pirate Bug
##                                   60                                          62
##                    European Dark Bee                                    Wireworm
##                                   66                                          69
##                       Euonymus Scale                           Asian Lady Beetle
##                                   75                                          76
##                      Japanese Beetle                            Italian Honeybee
##                                   94                                         113
##                           Bumble Bee                         Carniolan Honey Bee
##                                  140                                         152
##                  Buff Tailed Bumblebee                             Parasitic Wasp
##                                  183                                         285
##                            Honey Bee                                     (Other)
##                                  667                                         670
```

> Answer:The six most commonly studied species are the Italian honeybee, Bumble Bee, Carniolan Honey Bee, Buff Tailed Bumblebee, Parasitic Wasp, and Honey Bee (with "other" adding up to 670). All of these bees are very important pollinator species for agricultural crops, and plants of many kinds (flowers, trees, etc.) The parasitic wasps are also really good pollinators and are often released on farms to serve as natural biocontrol of insect pest populations.

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?

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

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

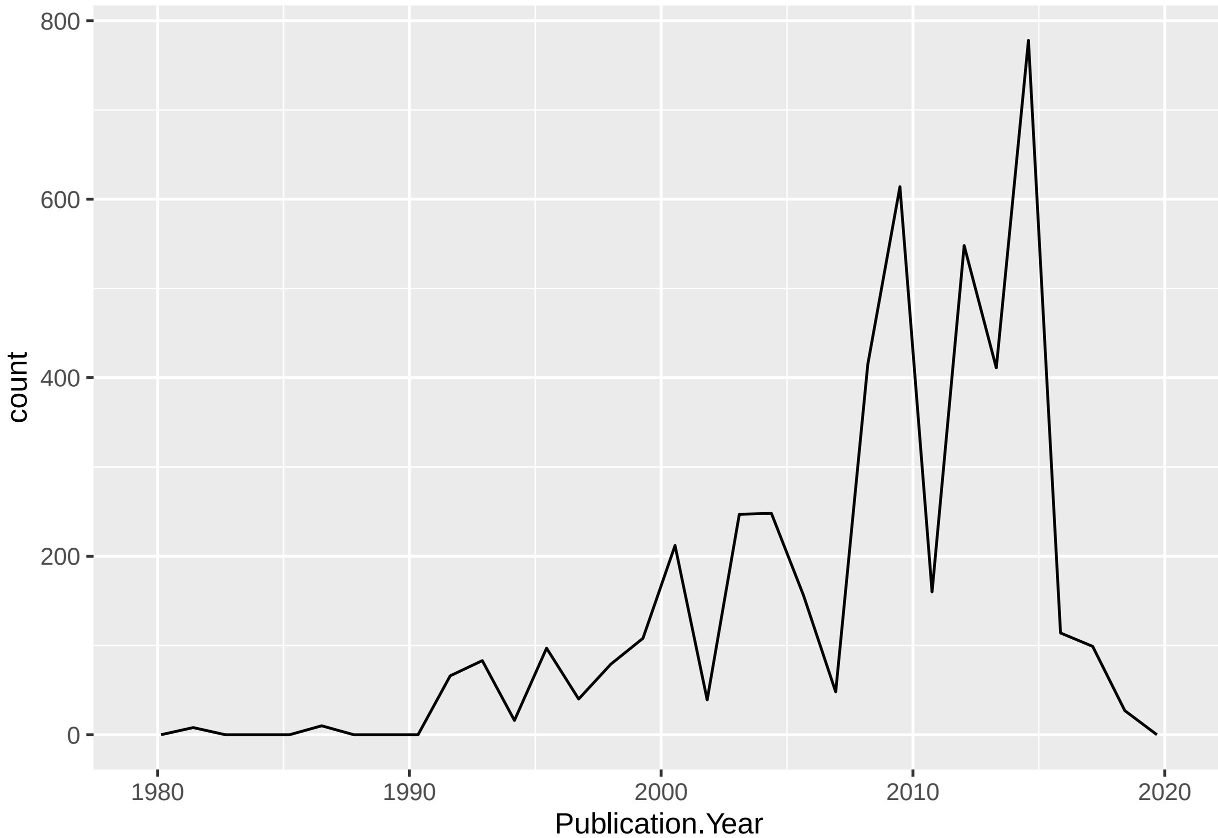
> Answer: It is a factor class, which is because we told R to read strings as factors, and so it is reading this whole column as a string, and thus, factors (in quotations).

Explore your data graphically (Neonics)

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

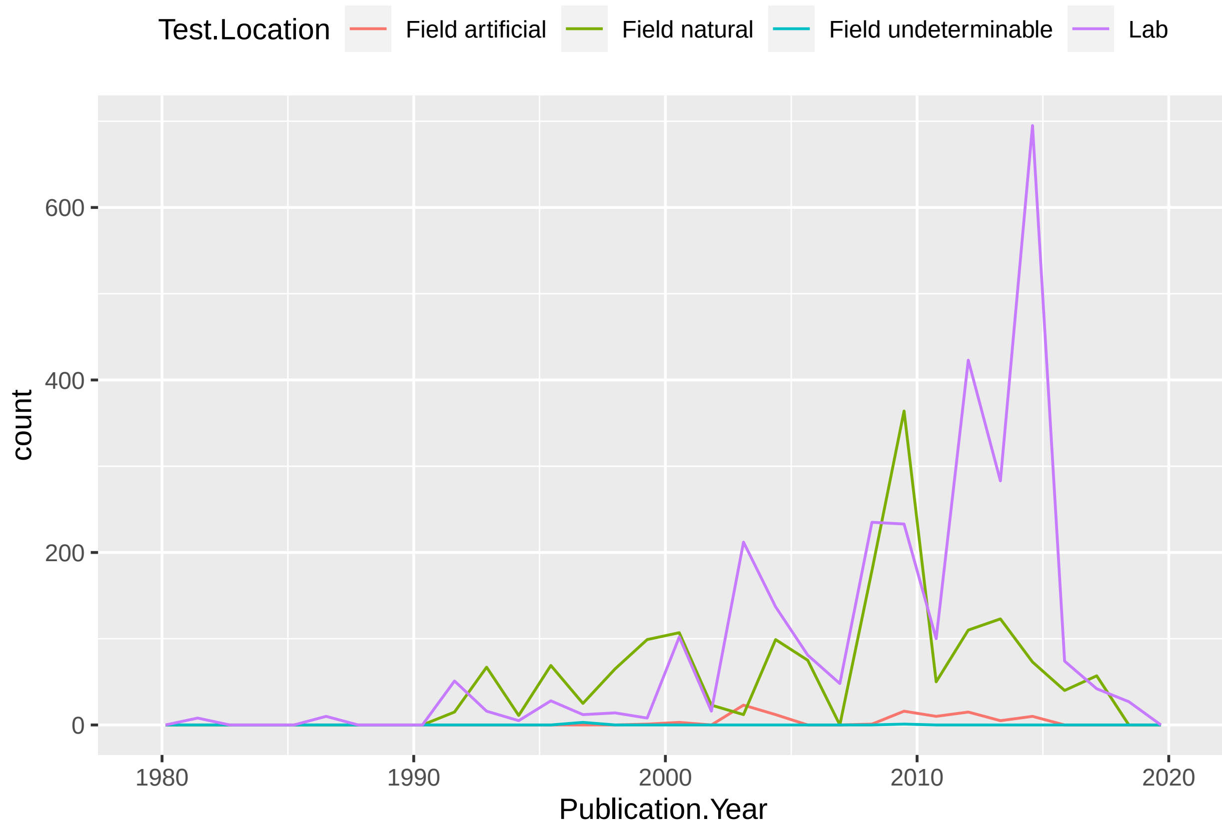
```
ggplot(Neonics, aes(Publication.Year)) +
geom_freqpoly()
```

```
## `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), bins = 30) +
theme(legend.position = "top")
```



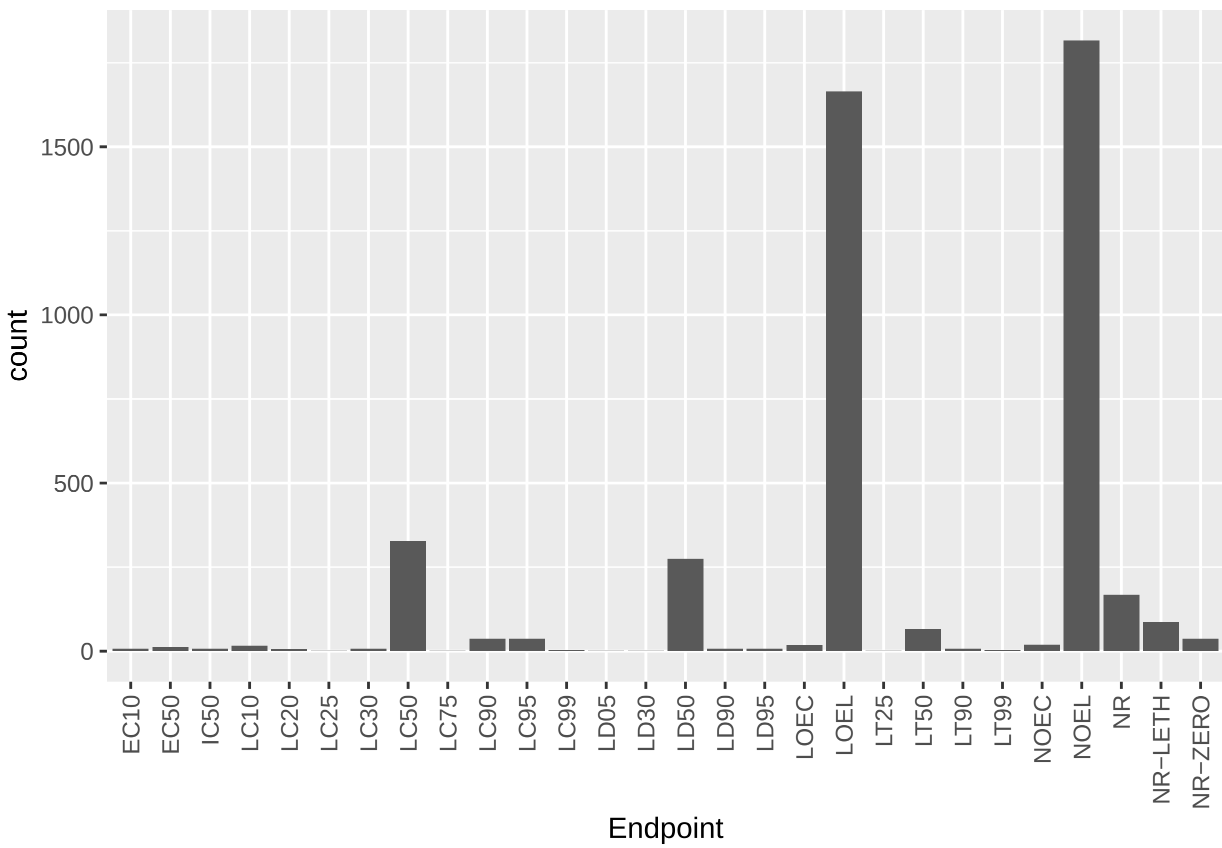
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The two most common test locations are the field natural and the lab. For many years, they were close to each other in how many samples were conducted in each, but around 2010, there was a dramatic increase in the number of samples done in the lab compared to ones done in the field. I wonder if this could be due to data that suggested neonics had negative impacts on the ecosystem, and thus testing them in the field became a non-viable option.

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**:`theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))` Add to the end of your plot command to rotate and align the X-axis labels…]

```
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 endpoints are LC50 and LOEL. LC50 stands for lethal concentration to xx% of test animals, for this endpoint it is 50%. LOEL stands for lowest-observable-effect-level, which was the "lowest dose producing effects that were significantly different (as reported by authors) from responses of controls" according to the Appendix.

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"
```

```
str_today <- Litter$collectDate
date_obj_today <- ymd(Litter$collectDate)
date_obj_today
```

```
##   [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"
```

```
class(date_obj_today)
```

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

```
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`?

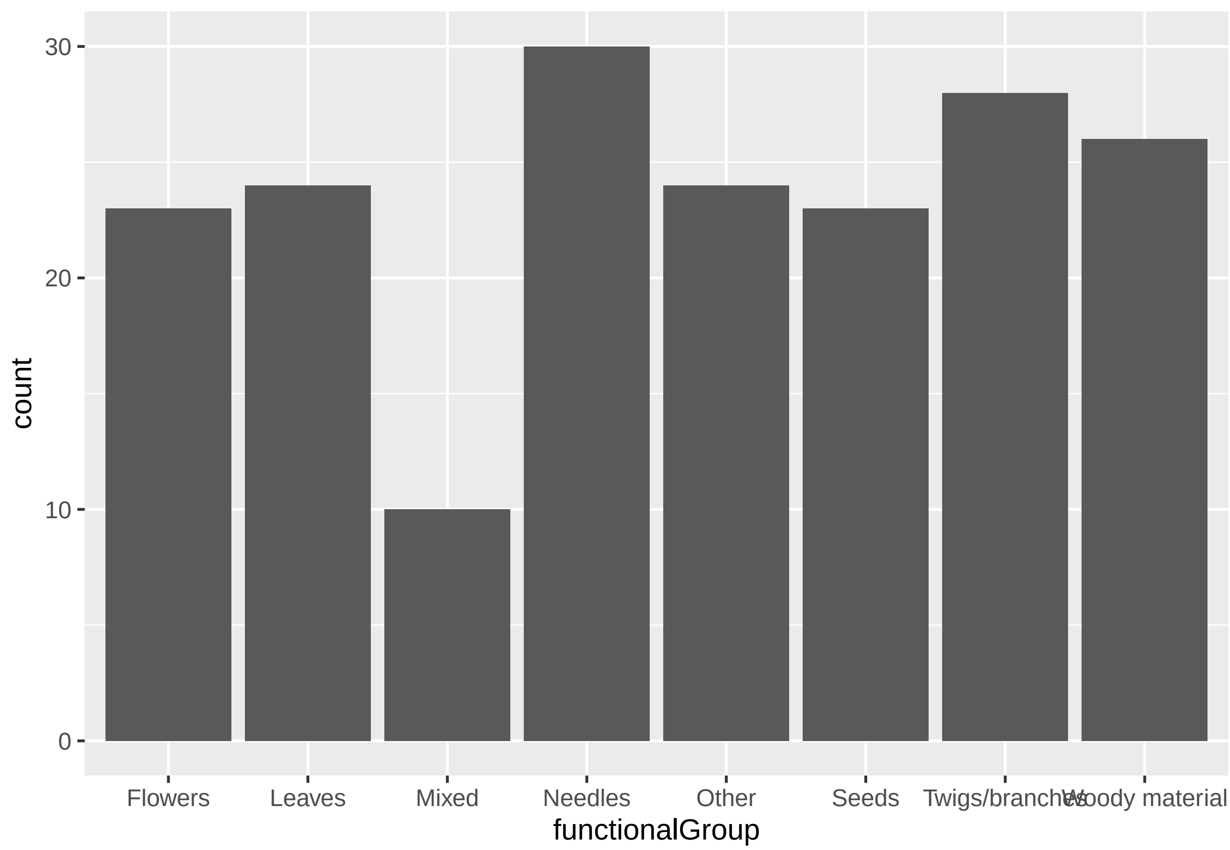
```
unique(Litter$plotID)
```

```
##  [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
```

Answer: 12 plots were sampled at Niwot Ridge. The unique function deletes/removes any duplicate data and does not include it in the output at all, whereas the summary would still include duplicate data.

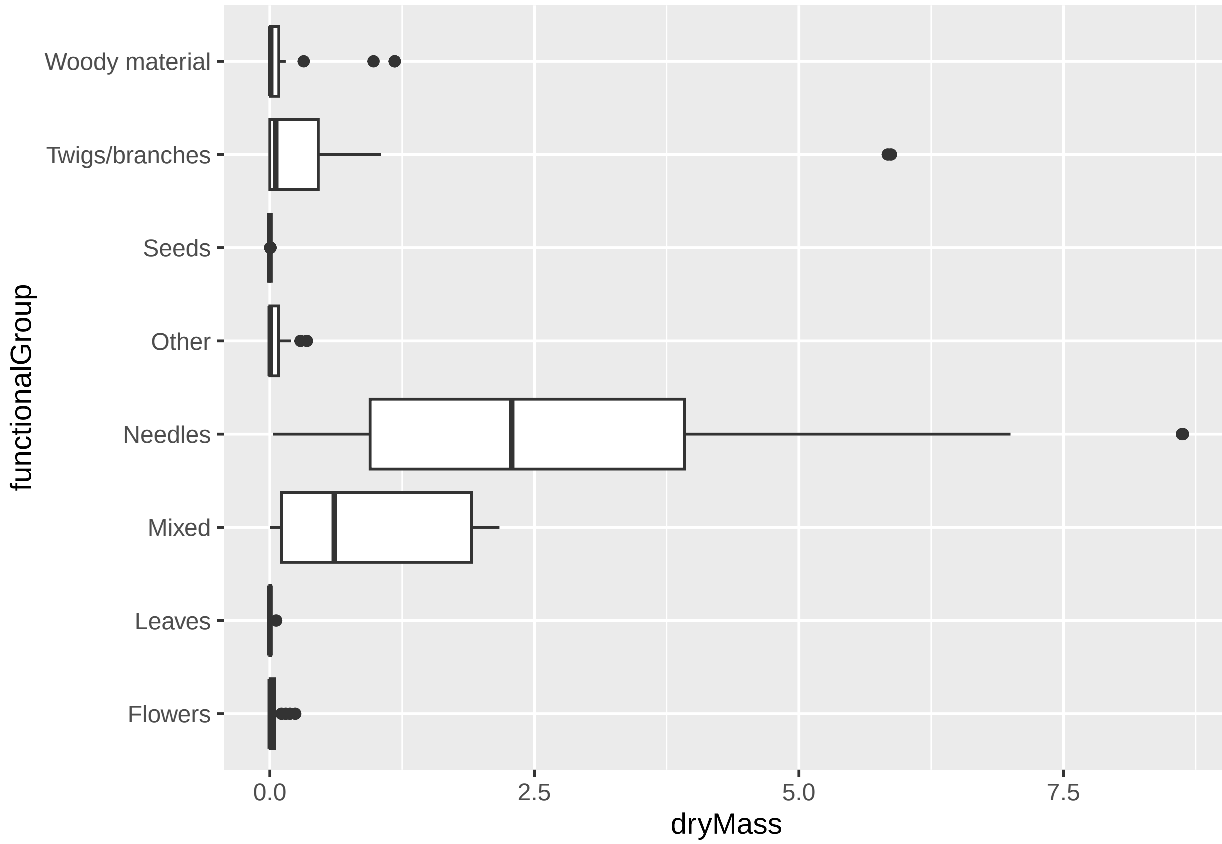
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

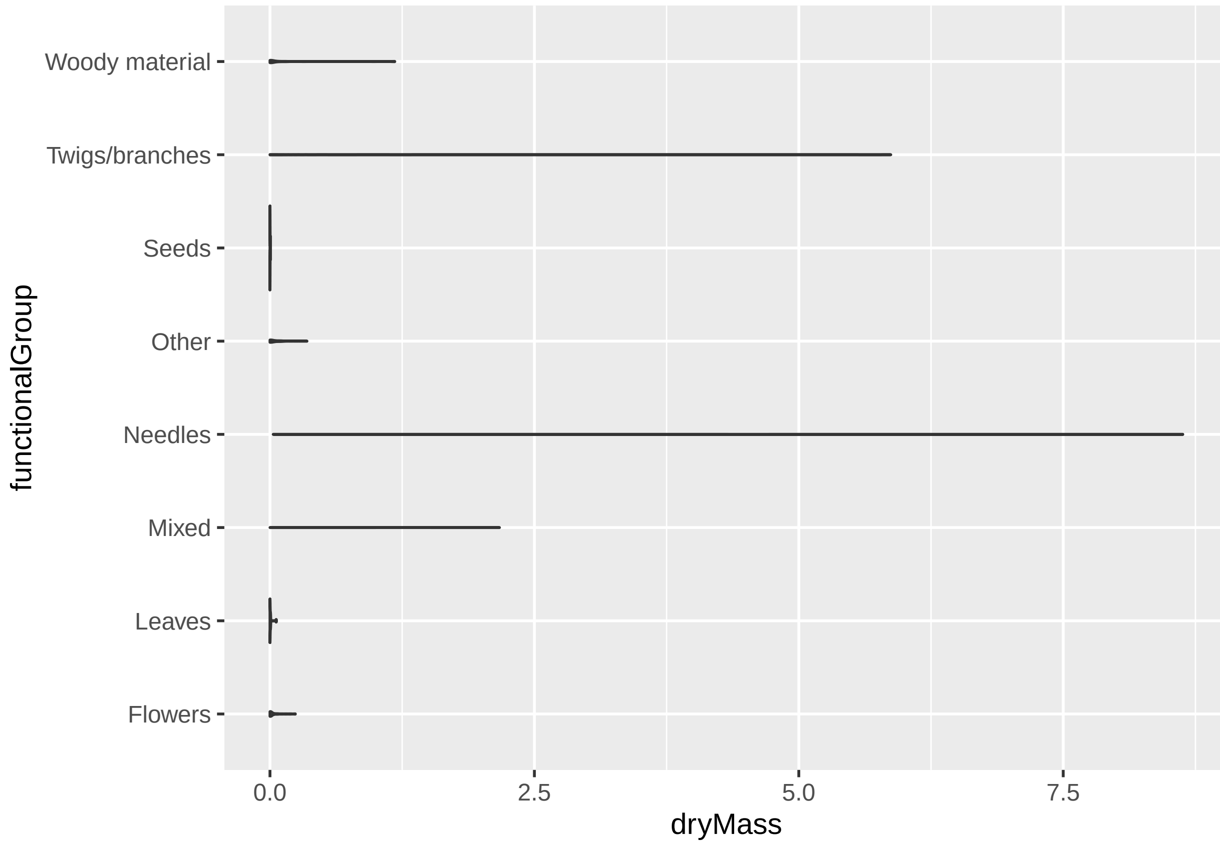


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




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



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

> Answer: Within this case, the boxplot depicts critical statistics like the confidence intervals/quantiles as well as median in a way that the violin plot does not capture. I don't know why, but the way the data distribution worked out in this dataset means the violin plot just looks like straight lines and you can't interpret much from those.

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

> Answer: Of the litter types, needles is the highest biomass, followed by twigs/branches.