Assignment 3: Data Exploration

John Rooney

Spring 2023

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 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)
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",stringsAsFactors = T)
Litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv",stringsAsFactors = T)
```

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 in insecticides have been linked to bee deaths, which is an issue of concern since bees pollinate approximately 80% of all flowering plants. I don't know about you, but I like food and would like it to stick around.

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: One reason we might be interested in studying litter and woody debris that falls to the ground in forests is that, left unmanaged, accumulated litter and woody debris poses a wildfire hazard as parts of the country continue to experience drought. Another is that I imagine they directly impact the kind of species that can thrive in a particular ecosystem.

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. Ground traps are only sampled once per year, with sampling frequency for elevated sites differing based on vegetation type. 2. Sites with decidious vegetation, or with limited access in winter months, may have sampling paused for up to six months. 3. Trap placement within plots may be targeted or randomized, meaning our sample is not completely random. This impacts what inferences we may be able to make.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

The Neonics dataset has 4623 observations of 30 variables.

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: The most common effects studied appear to be mortality and population. These are likely to be specifically of interest given the concern that neonicotinoids are causing a collapse of the bee population.

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), decreasing=T)
```

```
##                          (Other)                       Honey Bee
##                              670                             667
##                   Parasitic Wasp             Buff Tailed Bumblebee
##                              285                             183
##               Carniolan Honey Bee                      Bumble Bee
##                              152                             140
##                  Italian Honeybee                 Japanese Beetle
##                              113                              94
##                Asian Lady Beetle                  Euonymus Scale
##                               76                              75
##                         Wireworm                European Dark Bee
##                               69                              66
##                Minute Pirate Bug              Asian Citrus Psyllid
##                               62                              60
##                    Parastic Wasp            Colorado Potato Beetle
##                               58                              57
##                  Parasitoid Wasp               Erythrina Gall Wasp
##                               51                              49
##                     Beetle Order       Snout Beetle Family, Weevil
##                               47                              47
##          Sevenspotted Lady Beetle                  True Bug Order
##                               46                              45
##             Buff-tailed Bumblebee                    Aphid Family
##                               39                              38
##                   Cabbage Looper              Sweetpotato Whitefly
##                               38                              37
##                    Braconid Wasp                    Cotton Aphid
##                               33                              33
##                   Predatory Mite            Ladybird Beetle Family
##                               33                              30
##                      Parasitoid                    Scarab Beetle
##                               30                              29
##                   Spring Tiphia                     Thrip Order
##                               29                              29
##               Ground Beetle Family               Rove Beetle Family
##                               27                              27
##                   Tobacco Aphid                    Chalcid Wasp
```

```
##                              27                                  25
##             Convergent Lady Beetle                      Stingless Bee
##                              25                                  25
##                 Spider/Mite Class                 Tobacco Flea Beetle
##                              24                                  24
##                  Citrus Leafminer                    Ladybird Beetle
##                              23                                  23
##                         Mason Bee                           Mosquito
##                              22                                  22
##                     Argentine Ant                              Beetle
##                              21                                  21
##          Flatheaded Appletree Borer                Horned Oak Gall Wasp
##                              20                                  20
##                 Leaf Beetle Family                   Potato Leafhopper
##                              20                                  20
##          Tooth-necked Fungus Beetle                      Codling Moth
##                              20                                  19
##           Black-spotted Lady Beetle                      Calico Scale
##                              18                                  18
##                Fairyfly Parasitoid                       Lady Beetle
##                              18                                  18
##              Minute Parasitic Wasps                         Mirid Bug
##                              18                                  18
##                  Mulberry Pyralid                          Silkworm
##                              18                                  18
##                     Vedalia Beetle                Araneoid Spider Order
##                              18                                  17
##                          Bee Order                     Egg Parasitoid
##                              17                                  17
##                     Insect Class             Moth And Butterfly Order
##                              17                                  17
##       Oystershell Scale Parasitoid Hemlock Woolly Adelgid Lady Beetle
##                              17                                  16
##              Hemlock Wooly Adelgid                              Mite
##                              16                                  16
##                       Onion Thrip                Western Flower Thrips
##                              16                                  15
##                     Corn Earworm                  Green Peach Aphid
##                              14                                  14
##                         House Fly                           Ox Beetle
##                              14                                  14
##                 Red Scale Parasite                  Spined Soldier Bug
##                              14                                  14
##              Armoured Scale Family                  Diamondback Moth
##                              13                                  13
##                     Eulophid Wasp                  Monarch Butterfly
##                              13                                  13
##                     Predatory Bug                Yellow Fever Mosquito
##                              13                                  13
##                Braconid Parasitoid                       Common Thrip
##                              12                                  12
##       Eastern Subterranean Termite                             Jassid
##                              12                                  12
##                         Mite Order                          Pea Aphid
```

```
##                                12                                12
##                  Pond Wolf Spider          Spotless Ladybird Beetle
##                                12                                11
##             Glasshouse Potato Wasp                         Lacewing
##                                10                                10
##            Southern House Mosquito           Two Spotted Lady Beetle
##                                10                                10
##                         Ant Family                      Apple Maggot
##                                 9                                 9
```

Answer: The six most commonly studied species in the dataset (common name) are honey bees, parasitic wasps, buff tailed bumblebees, carniolan honey bees, bumble bees, and Italian honeybees. With the exception of parasitic wasps, I would suggest these are of interest more than the others because of the link between neonicotinoids and bee deaths.

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.)
```

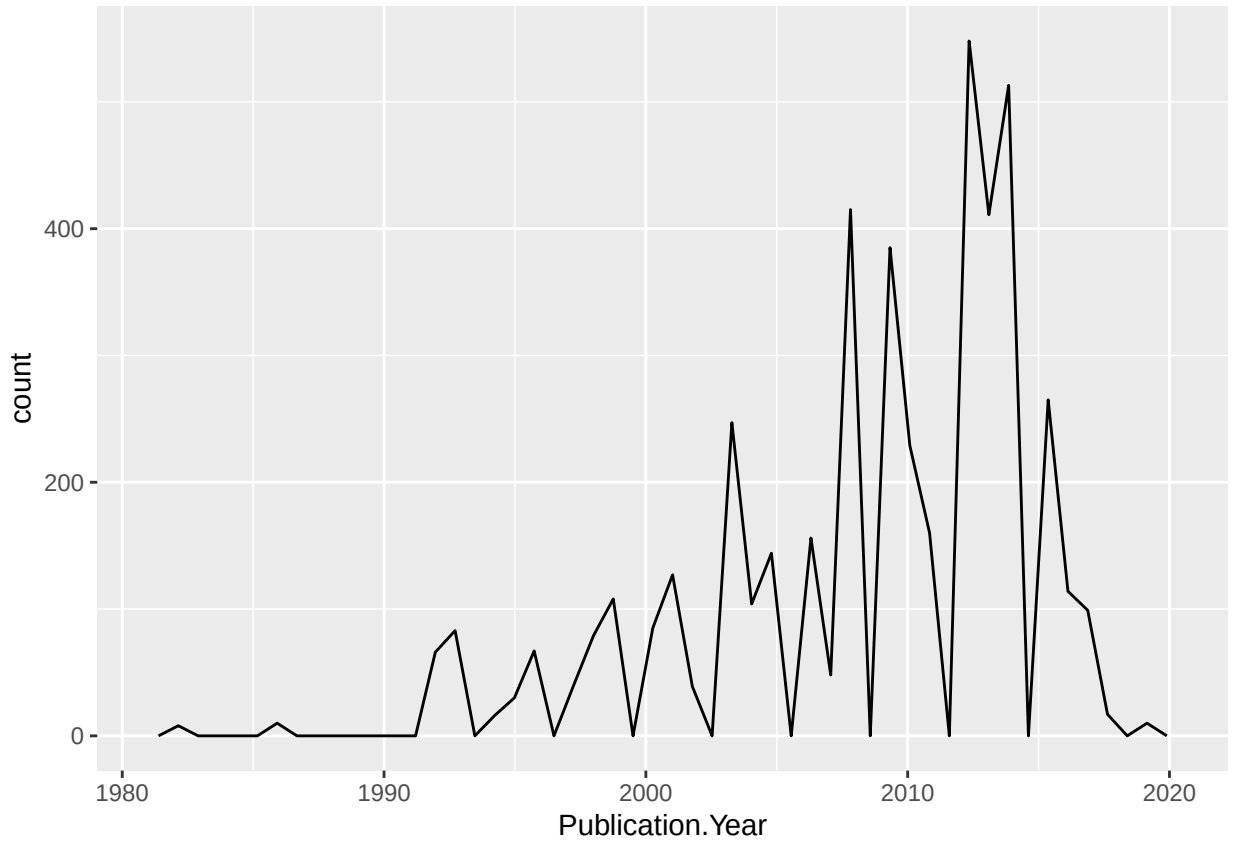
```
## [1] "factor"
```

Answer: The class of "Conc.1..Author" is a factor. The values in that column are "Active Ingredient", "Formulation", and "Not Coded". Because these values are words that R would read as in quotations when referenced, they're considered strings, and we've told R to read strings as factors.

Explore your data graphically (Neonics)

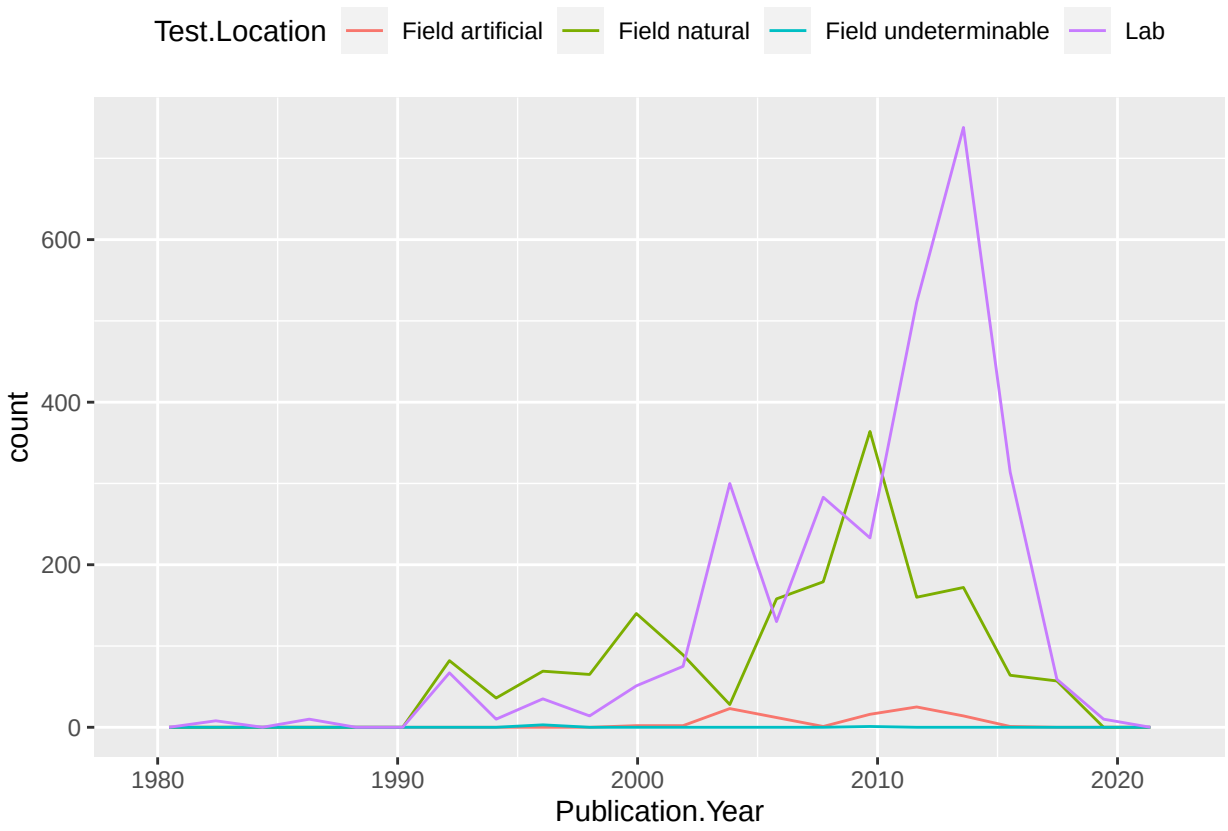
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50)
```



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=20)+
  theme(legend.position = "top")
```



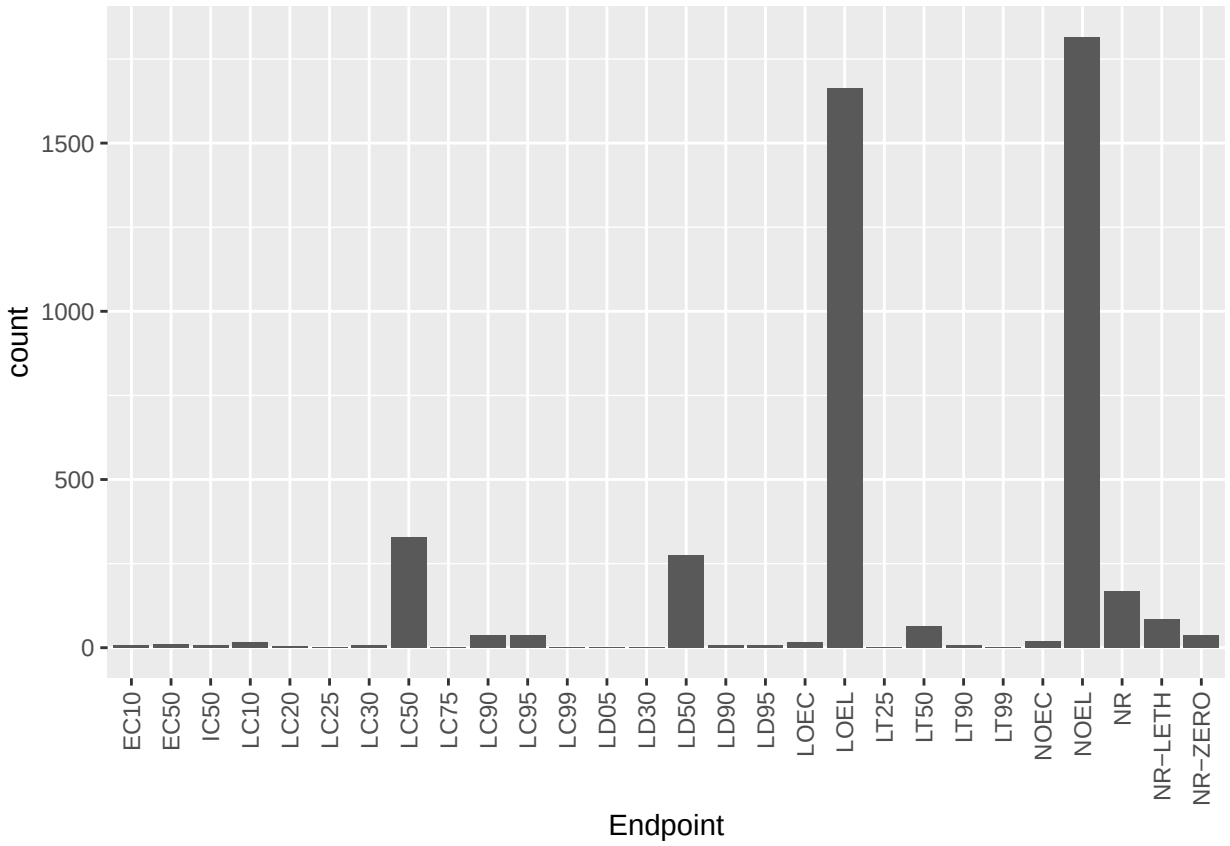
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The two most common test locations are lab and field natural. During the 90s they were relatively close in count, where more location sites in the field. This shifts in what looks like 2004, where field sites drop and lab sites increase. The switch places through the end of the decade, when the count for lab test locations rises dramatically and field sites plummet.

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: The two most common end points are LOEL and NOEL, which stand for Lowest Observable Effect Level and No Observable Effect Level. The former tells us the lowest dose (or concentration) of neonics that produce effects significantly different than those in the control group, and the latter tells us the highest dose (or concentration) of neonics that do *not* produce effects significantly different than those for in 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"
```

```
str_collectDate <- Litter$collectDate
date_obj_today <- ymd(str_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(date_obj_today)
```

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

The initial class for collectData is factor. The dates litter was sampled are August 2 and August 30, 2018.

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$namedLocation)
```

```
##  [1] NIWO_061.basePlot.ltr NIWO_064.basePlot.ltr NIWO_067.basePlot.ltr
##  [4] NIWO_040.basePlot.ltr NIWO_041.basePlot.ltr NIWO_063.basePlot.ltr
```

```
##  [7] NIWO_047.basePlot.ltr NIWO_051.basePlot.ltr NIWO_058.basePlot.ltr
## [10] NIWO_046.basePlot.ltr NIWO_062.basePlot.ltr NIWO_057.basePlot.ltr
## 12 Levels: NIWO_040.basePlot.ltr ... NIWO_067.basePlot.ltr
```

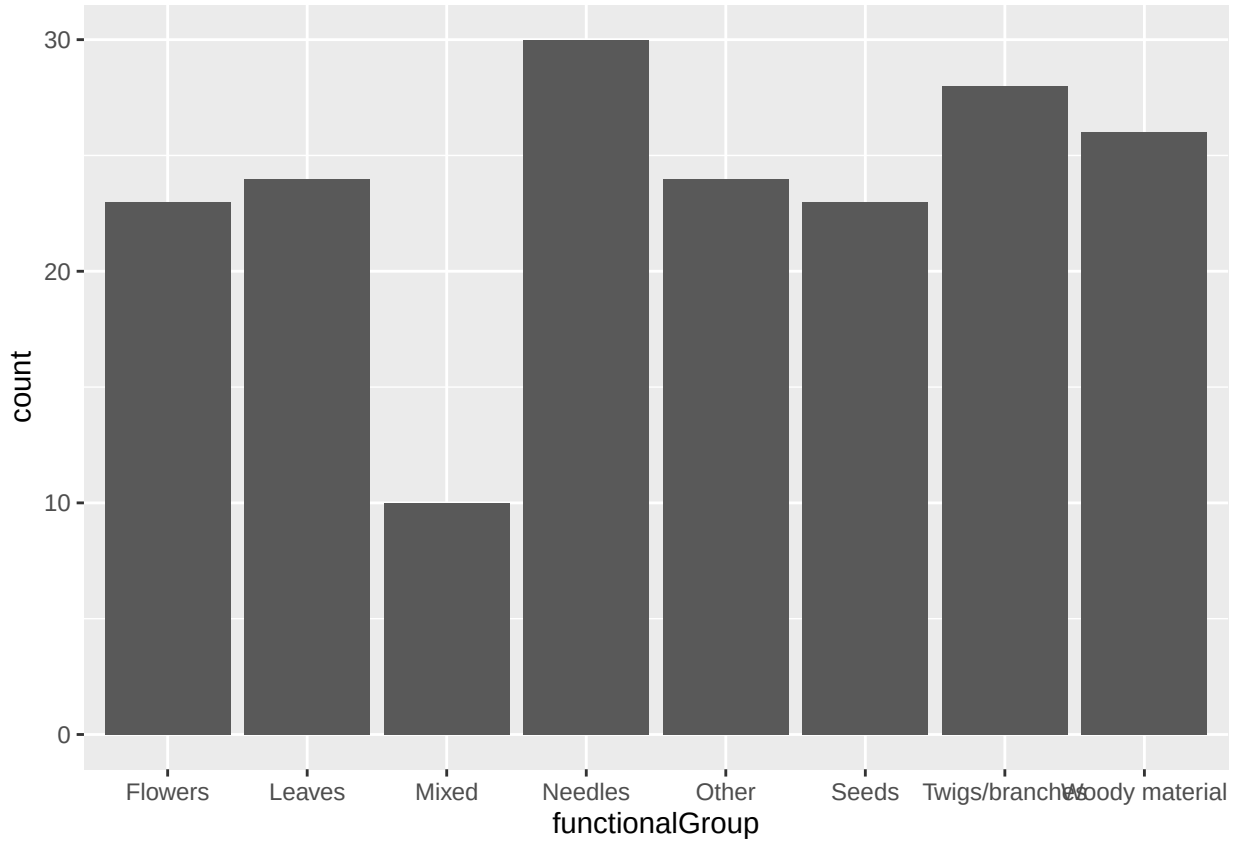
```
summary(Litter$namedLocation)
```

```
## NIWO_040.basePlot.ltr NIWO_041.basePlot.ltr NIWO_046.basePlot.ltr
##                    20                    19                    18
## NIWO_047.basePlot.ltr NIWO_051.basePlot.ltr NIWO_057.basePlot.ltr
##                    15                    14                     8
## NIWO_058.basePlot.ltr NIWO_061.basePlot.ltr NIWO_062.basePlot.ltr
##                    16                    17                    14
## NIWO_063.basePlot.ltr NIWO_064.basePlot.ltr NIWO_067.basePlot.ltr
##                    14                    16                    17
```

Answer:There were 12 different plots sampled at Niwot Ridge. Using "summary" provided a count of how many samples were collected at each site, while "unique" simply listed out each unique site without telling us how many samples were taken at each site.

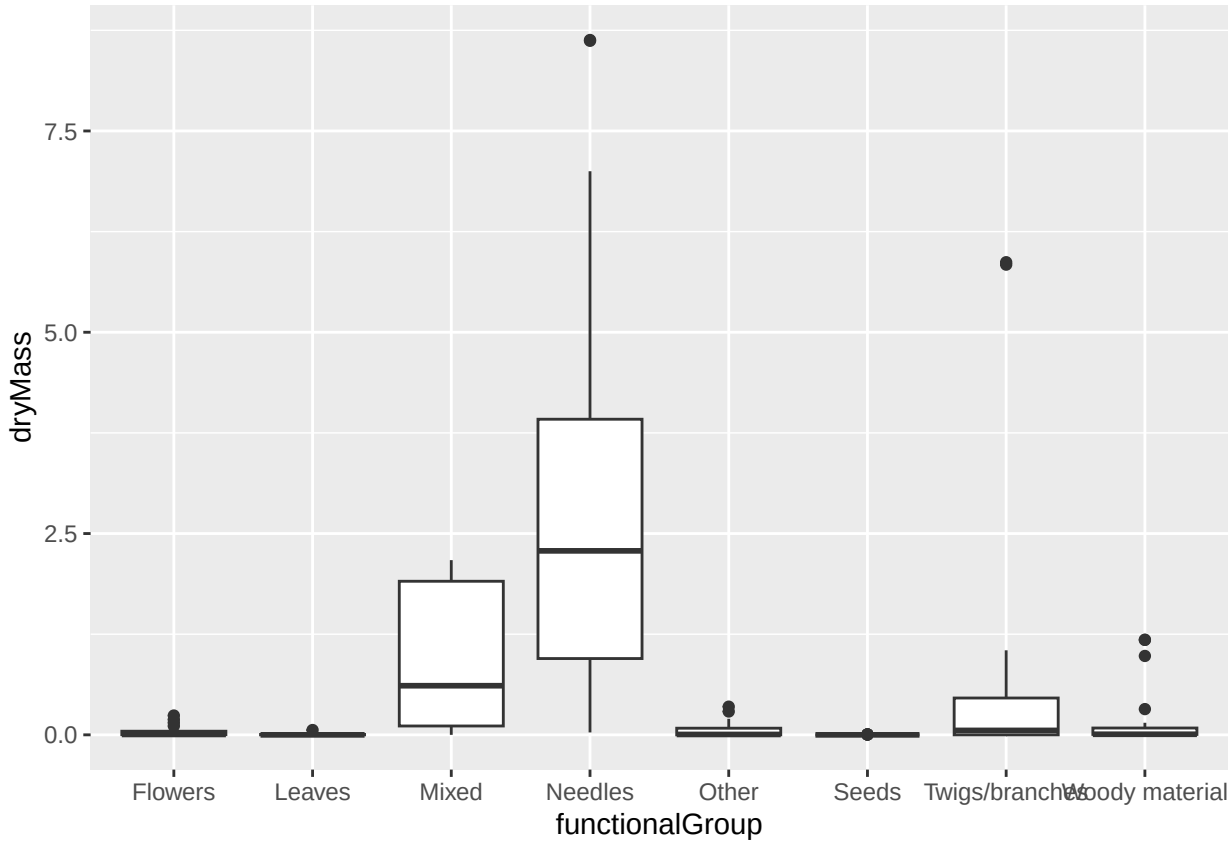
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

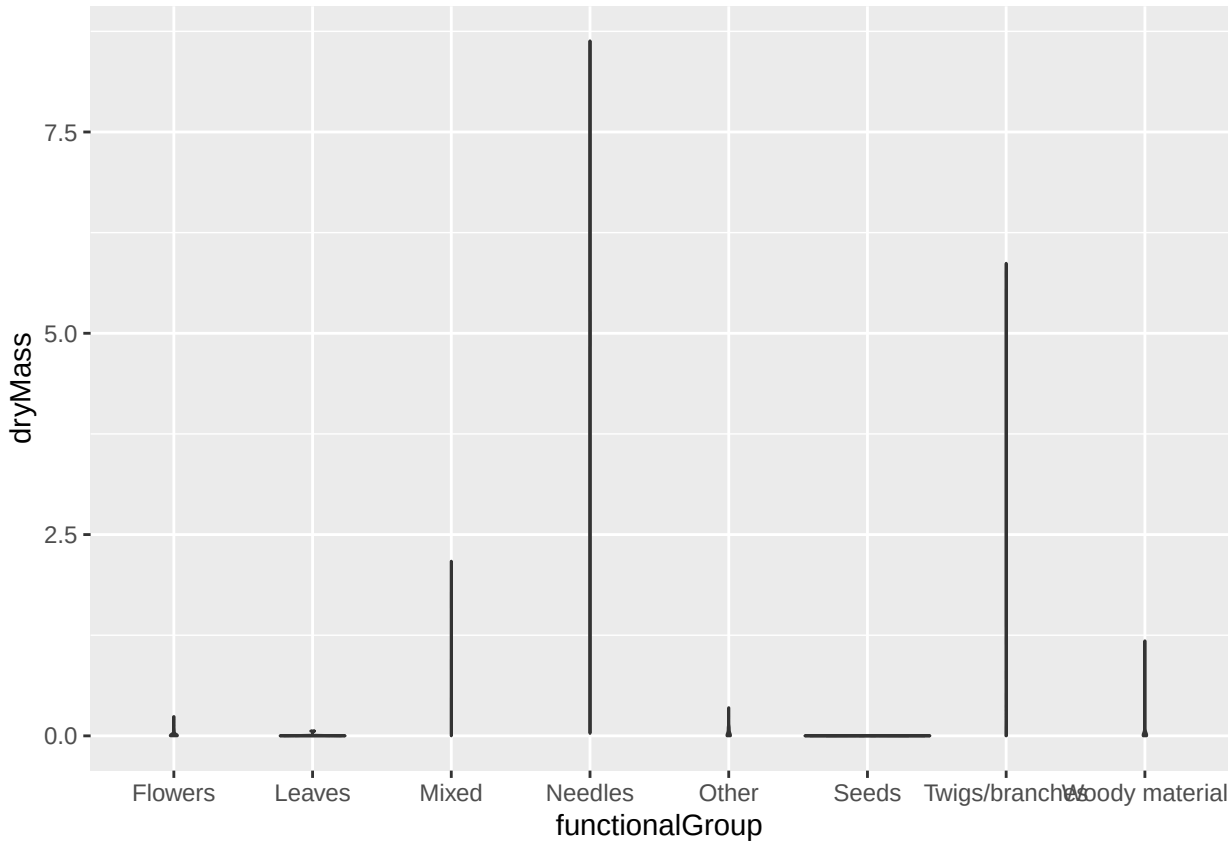


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




```
ggplot(Litter)+
  geom_violin(aes(x = functionalGroup, y = dryMass),
              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: Violin plots are new to me, and I'm not quite sure why my violin plot isn't showing like it did in the example markdown (ie. with more shape). The only ones that aren't appearing as lines are leaves and seeds, which show some width at the bottom. In this case we are getting far more data from the boxplot as it is showing summary statistics (min, max, median, IRQ) and outliers.

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

> Answer: Needles have far and away the highest biomass at these sites, followed by mixed litter.