Assignment 3: Data Exploration

Cal Oakley, Section #2

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Exploration.

Directions

1. Change "Student Name, Section #" on line 3 (above) with your name and section number.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "FirstLast_A03_DataExploration.Rmd") prior to submission.

The completed exercise is due on <>.

Set up your R session

1. Check your working directory, load necessary packages (tidyverse), 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 add the `stringsAsFactors = TRUE` parameter to the function when reading in the CSV files.**

```
setwd("K:/GradSchool/Spring2022/EnvironmentalDataAnalytics/Environmental_Data_Analytics_2022/Data/Raw")
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.2

## Warning: package 'ggplot2' was built under R version 4.1.2

## Warning: package 'tidyr' was built under R version 4.1.1

## Warning: package 'readr' was built under R version 4.1.1

## Warning: package 'dplyr' was built under R version 4.1.2

## Warning: package 'stringr' was built under R version 4.1.2

## Warning: package 'forcats' was built under R version 4.1.2
```

```
Neonics <- read.csv("ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = TRUE)
Litter <- read.csv("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 ecotoxicologoy of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

Answer: Neonicotinoids are manufactured chemicals that are similar to the natural compound nicotine. They are generally considered to be more harmful to insects than mammals. Since they are often used to control insects that attack crops, they tend to be broadcast into the environment in a manner that could expose non-target insect species which as implications for insect populations beyond the agro-ecosystem, especially if the chemicals are presist for extended periods.

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: Sampling forest litter and woody debris is important for a variety of reasons. It can be used to study nutrient cycling within forest ecosystems, assess wildfire hazard potential, and estimate carbon storage in soil.

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: It is sampled using elevated and ground traps to intercept litter and woody debris on its way from the canopy to the forest floor. Sampling are taken at terrestrial NEON sites with woody vegetation greater than 2 meters in height.
   Different plant organs (leaves, needles, twigs/branches, cones, flowers, seeds, fruit, bark, etc.) are weighed seperately to determine the mass of each group. Elevated and ground traps are different sizes and are designed to capture different types of plant organs. *All samples are oven dried before being weighed by technicians.

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?

```
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: From looking at these data, I think the most common effects are ‘Population’, ‘Mortality’, ‘Behavior’, ‘Feeding behavior’, ‘Reproduction’, and ‘Development’. These effects could be of interest because, as is evident from the names alone, they all relate to population dynamics. Some very directly, like ‘Population’, ‘Mortality’, & ‘Reproduction’; and other more indirectly, like ‘Behavior’, ‘Feeding behavior’, & ‘Development’. These can all be used to observe the scale and degree to which pesticide spillover from agriculture effects non-target insect species populations (the ‘direct’ effects) and the casual pathways involved (the ‘indirect’ effects).

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.

```
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 top six most commonly studied species are all bees (with the exception of 'Parasitic Wasp', which to me looks like a polyphyletic group; but still, close enough). These species are of particular interest relative to other insects because of their value in crop production (pollination -> plant reproduction -> fruit+veg). In agro-ecosystems there can be just as many desireable insects as there are undesireable ones. So, if a farmer is going to broadcast insecticides over their farm to control an undesireable species they need to be assured that it won't have a negative impact on the species they need to generate enough of a harvest.

8. Concentrations are always a numeric value. What is the class of Conc.1..Author. in the dataset, and why is it not numeric?

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

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

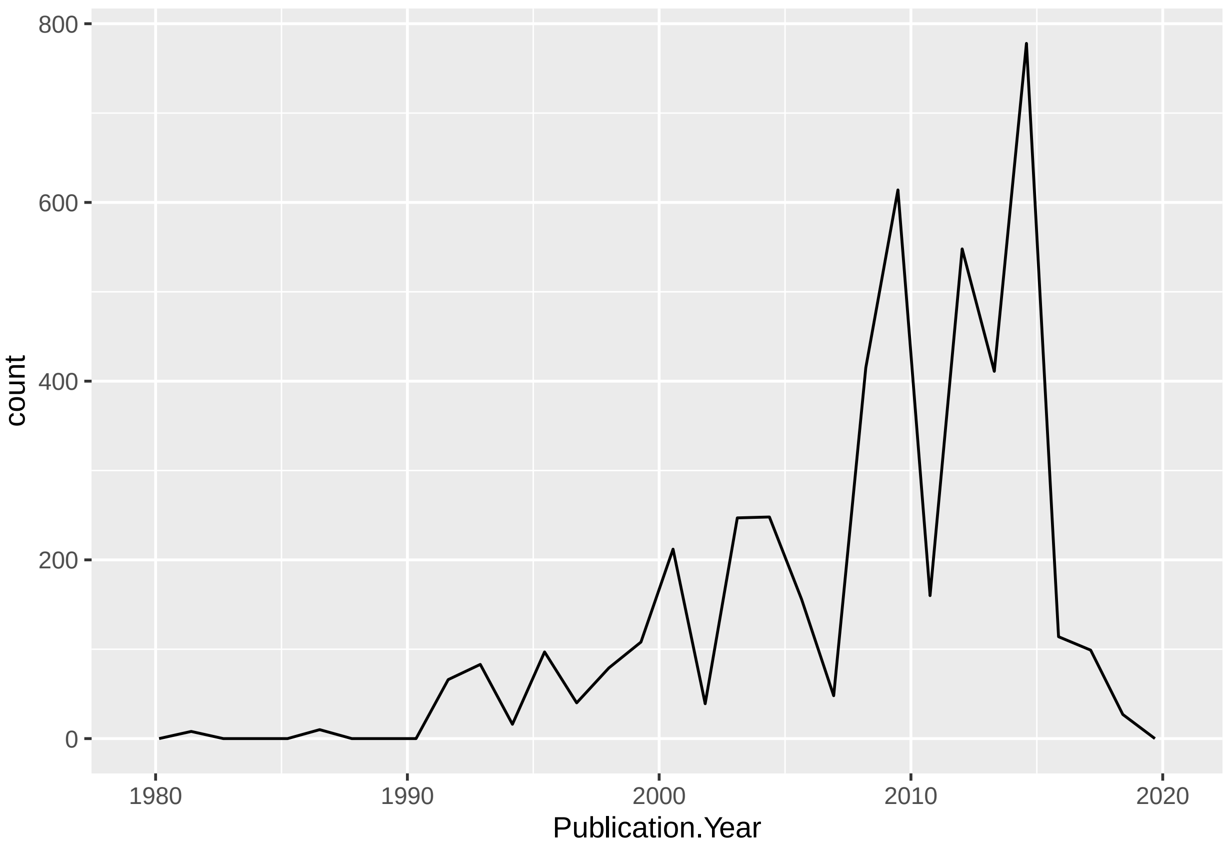
> Answer: 'Conc.1..Author.' is classed as "factor" in this dataset. I think what has happened here is that the values, although numeric, were classed as "string" in the raw CSV, so when I imported the dataset with the 'stringsAsFactors = TRUE' they were convereted to "factor".

Explore your data graphically (Neonics)

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

```
library(ggplot2)
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year))
```

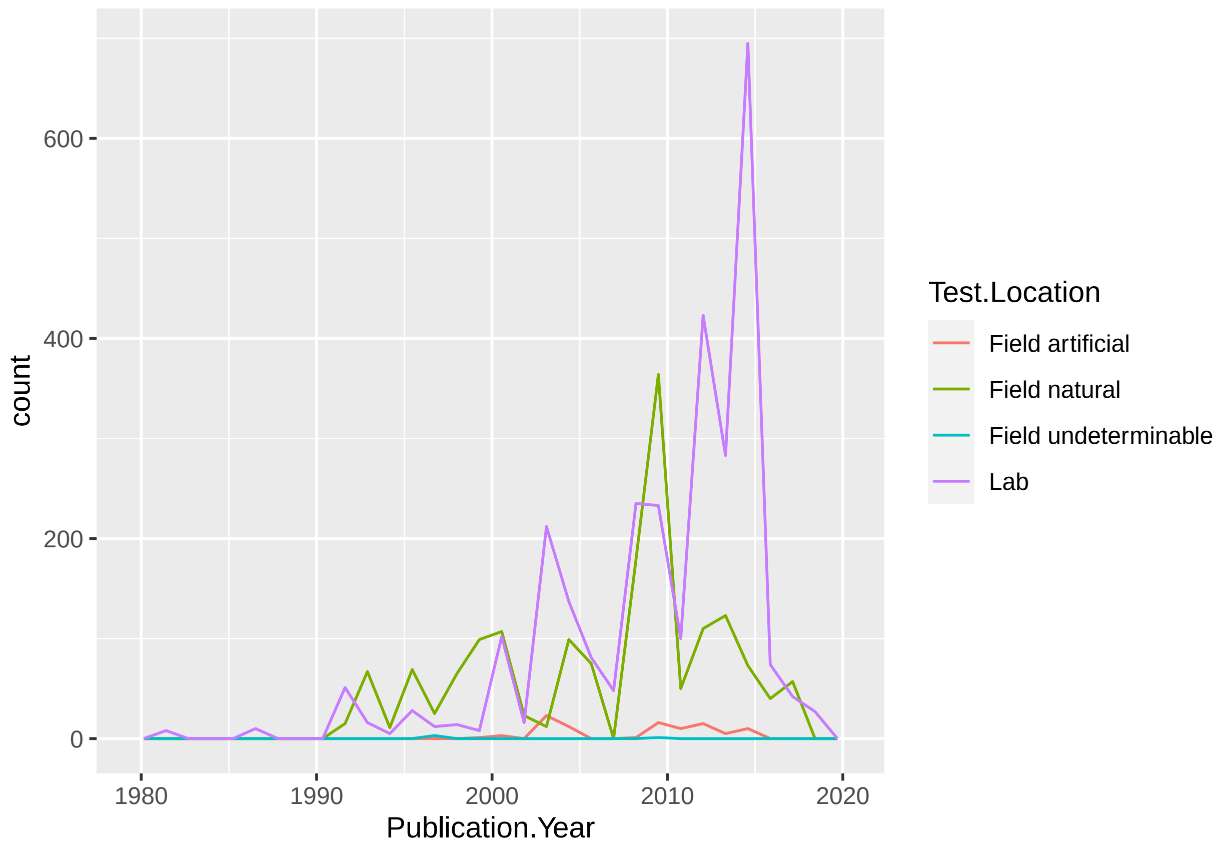
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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`.
```

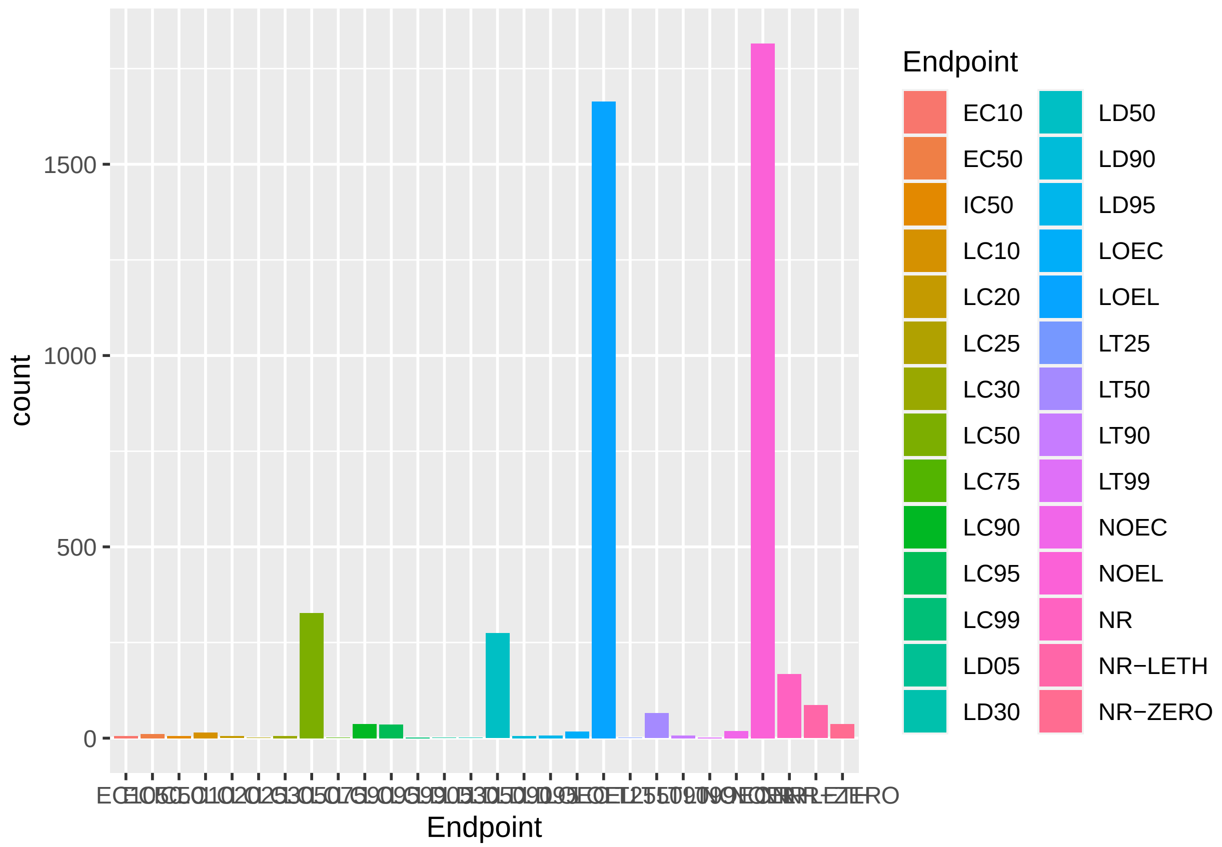


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are 'Lab' & 'Field natural'. The number of studies classified as 'Field natural' grew steadily between 1900 and the early 2000's and saw a peak around 2009-2010, but have since fallen. The number of studies classified as 'Lab' also began growing after 1990 albeit at a faster rate than 'Field natural'. 'Lab' studies peaked around 2015, but have also dropped off since then.

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.

```r
Neonics_df <- data.frame(Neonics)
ggplot(Neonics) +
  geom_bar(aes(x = Endpoint, fill = Endpoint))
```



Answer: The two most common endpoints are 'NOEL' and 'LOEL'. According to 'ECOTOX_CodeAppendix', 'NOEL' represents "no-observable-effect-level" which means the study reported the highest dose at which the insecticide in question produced no effects that differed significantly from control results. In other words, these studies reported a safe dose limit for the insecticide the were studying. 'LOEL' represents "Lowest Observed Effects Residue" which means the study reported the smallest dose at which observers found a response that differed significantly from the control.

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"
```

```
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`?

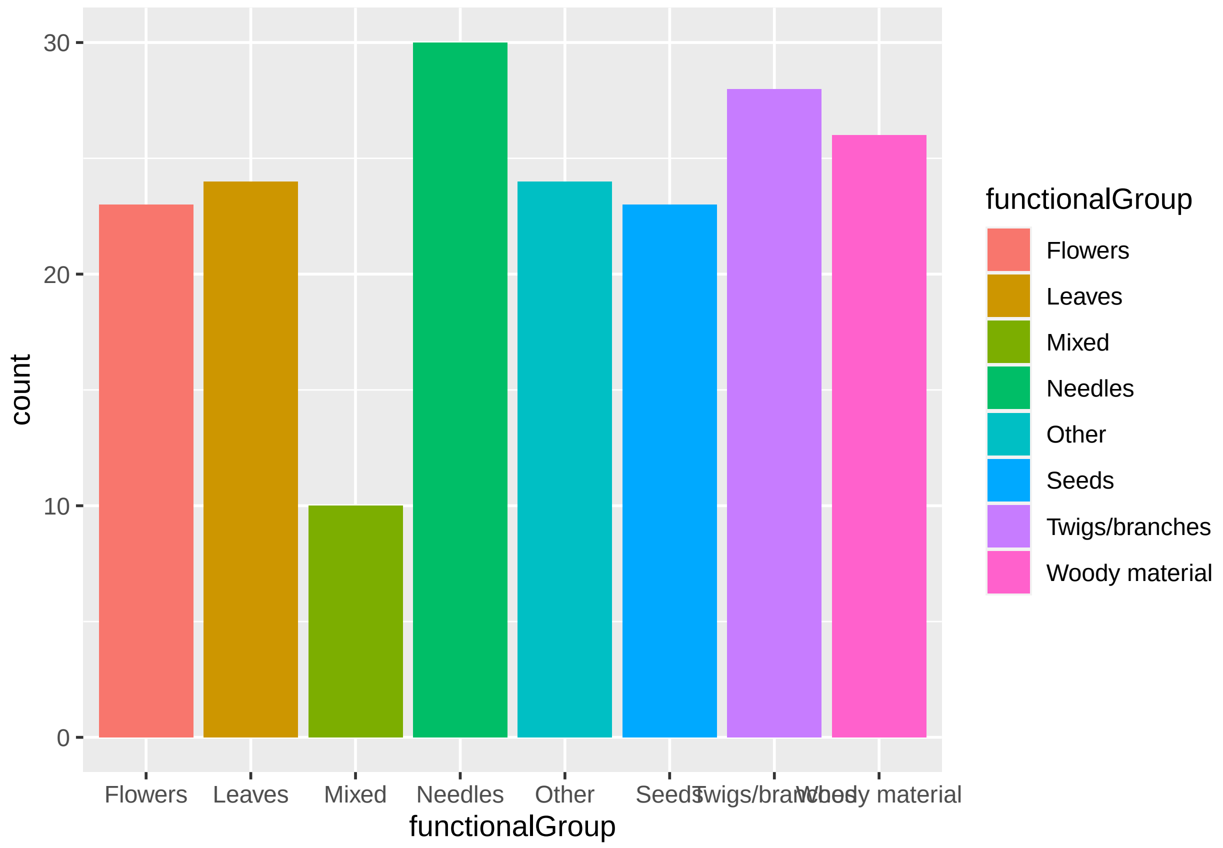
```
unique(Litter$plotID)
```

```
##  [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: 'summary' differs from 'unique' in that 'summary' looks at a field and returns the count of each unique type of answer, while 'unique' just returns a count of of those unique answer types.

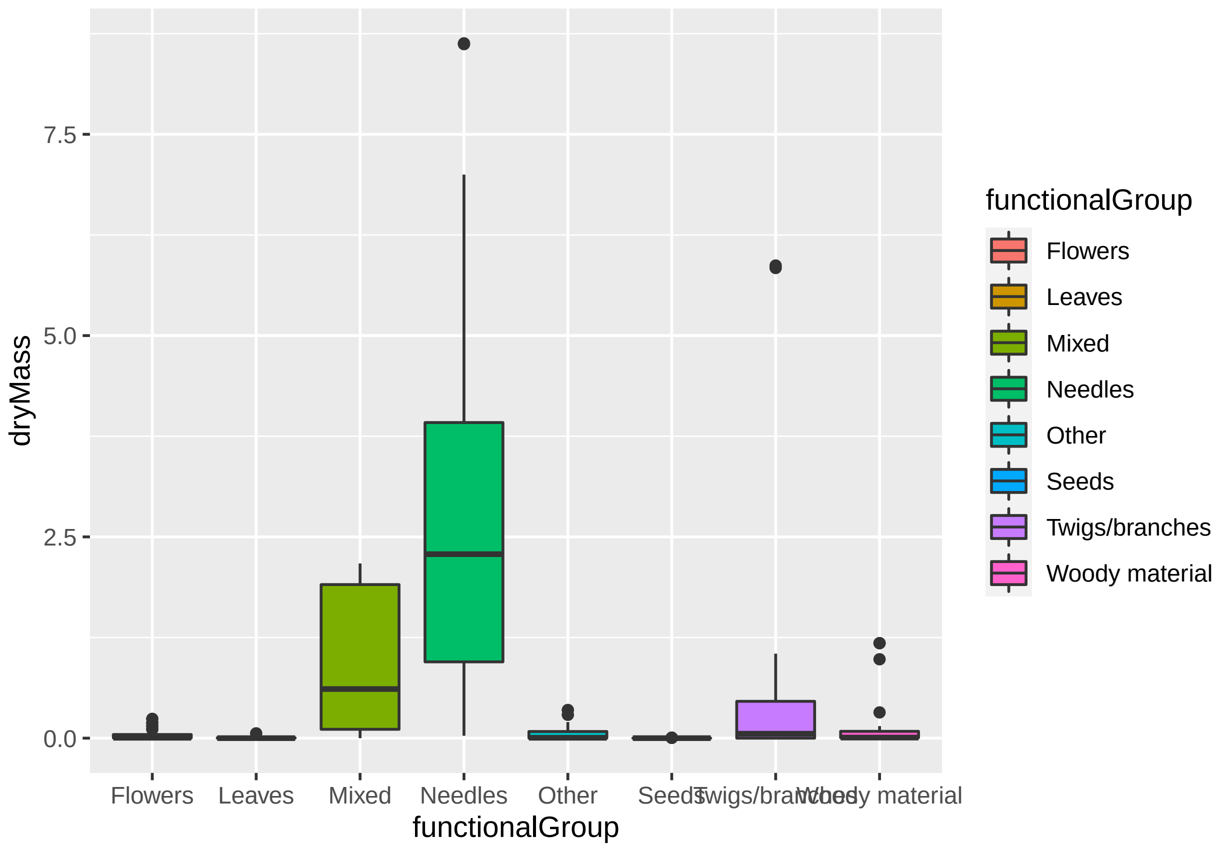
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

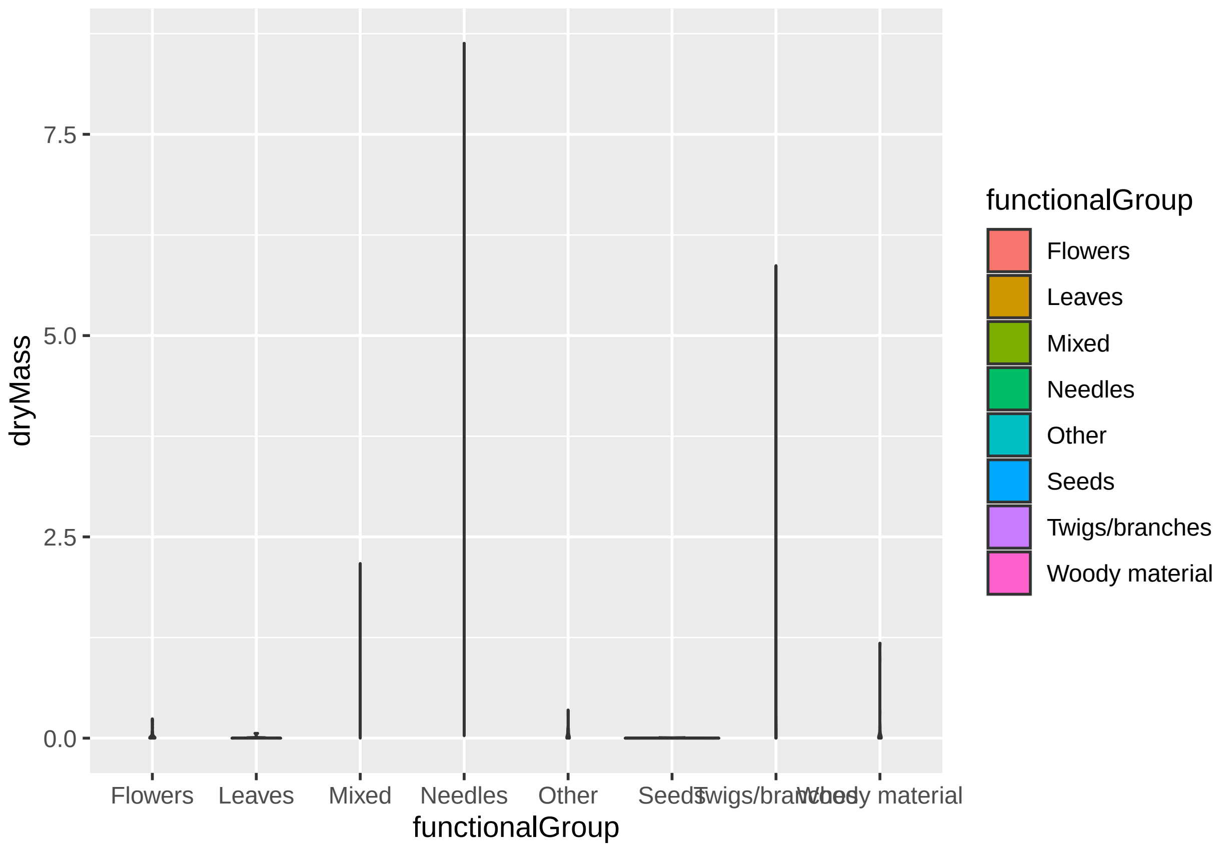


```
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass, fill = functionalGroup),
                draw_quantiles = c(0.25, 0.5, 0.75))
```

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```



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

> Answer: In this case, a boxplot is a more effective visualization option, compared to the violin plot, because the distribution of 'dryMass' within each 'functionalGroup' differs too much between each 'functionalGroups'. For example, 'Needles' 'dryMass' ranges between 0.0 and 7.5, with samples appearing to occur in similar, low frequencies throughout said range. Meanwhile, 'Seeds' 'dryMass' ranges between 0.0 and something just above that, with all samples occuring there. By asking ggplot() to put even just those two 'functionalGroups' on the same figure obfuscates all patterns in the data.

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

> Answer: At this site, 'Needles' and 'Mixed' litters tend to have the greatest biomass.