Assignment 3: Data Exploration

Katryna Niva

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. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.
6. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai.

The completed exercise is due on Sept 30th.

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 include the subcommand to read strings in as factors.

```
#install.packages('formatR')
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=80), tidy=TRUE)

library(tidyverse)
library(ggplot2)

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

Litter <- read.csv("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: Pesticides are having a large impact on endemic insect populations which can have dramatic long term impacts on ecosystem dynamics. Understanding how a given type of Neonicotinoids impacts a given type of insect can inform which pesticides can be used in which area without longterm consequences and which are best at targeting a specific pest.

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: The buildup of this woody debris has a large impact on fire hazard. The more dry, dead matter is there, the larger a fire can grow and should there be enough, a fire can grow hot enough that the live trees can catch and at that point, a fire spreading through the canopy is far more destructive.

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. The depris is collected in ground traps (3mx0.5m rectangles). Elevated traps (.5 m2 basket) are not used exclusively because they fail to reliably collect longer debris. 2. There are up to 12 litterfall collections per year with litter collected from up to 40 ground traps/sites each time. 3. Litter is defined as material that is dropped from the forest canopy and has a butt end diameter <2cm and a length <50cm.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
print(dim(Neonics))
```

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

```
# The dimensions of the Neonics dataset is 4623x30.
```

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?

```
print(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 include mortality and population. This is likely the case because changes in rate of mortality and changes in populations most directly reflect how the application of a pesticide will impact a given insect population. Whether looking to get rid of an insect that is eating crops and impacting yield or looking at a polinator that needs to be protected in order for successful harvests, these two points are the most important.

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.

```
print(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 studied species are the Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. All of these species are important polinators. Strong populations of polinators are imperative for most crops and so it makes sense that it is especially important for farmers to monitor the effect the pesticides they use have on polinators. Using a pesticide that kills off their polinators would negate all their efforts.

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

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

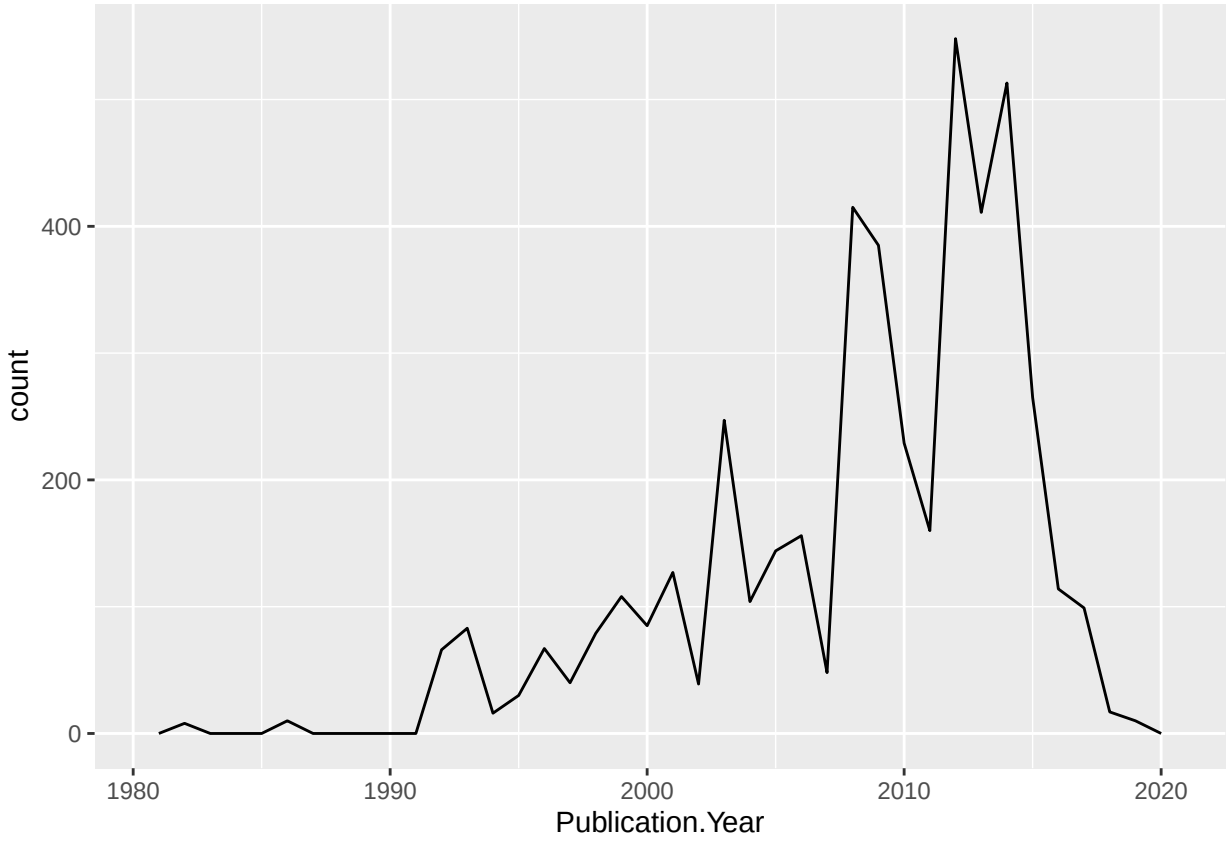
```
## [1] "factor"
```

Answer:The class of Conc.1..Author is factor. This is because certain inputs are NR (not reported) and since NA was not used for this, the code does not know to understand it as an empty point and so it is interpretting it as a character. This converts the entire column to character and we have told the code to consider character vectors as factors so that repeated messages (such as NR) can be grouped together.

Explore your data graphically (Neonics)

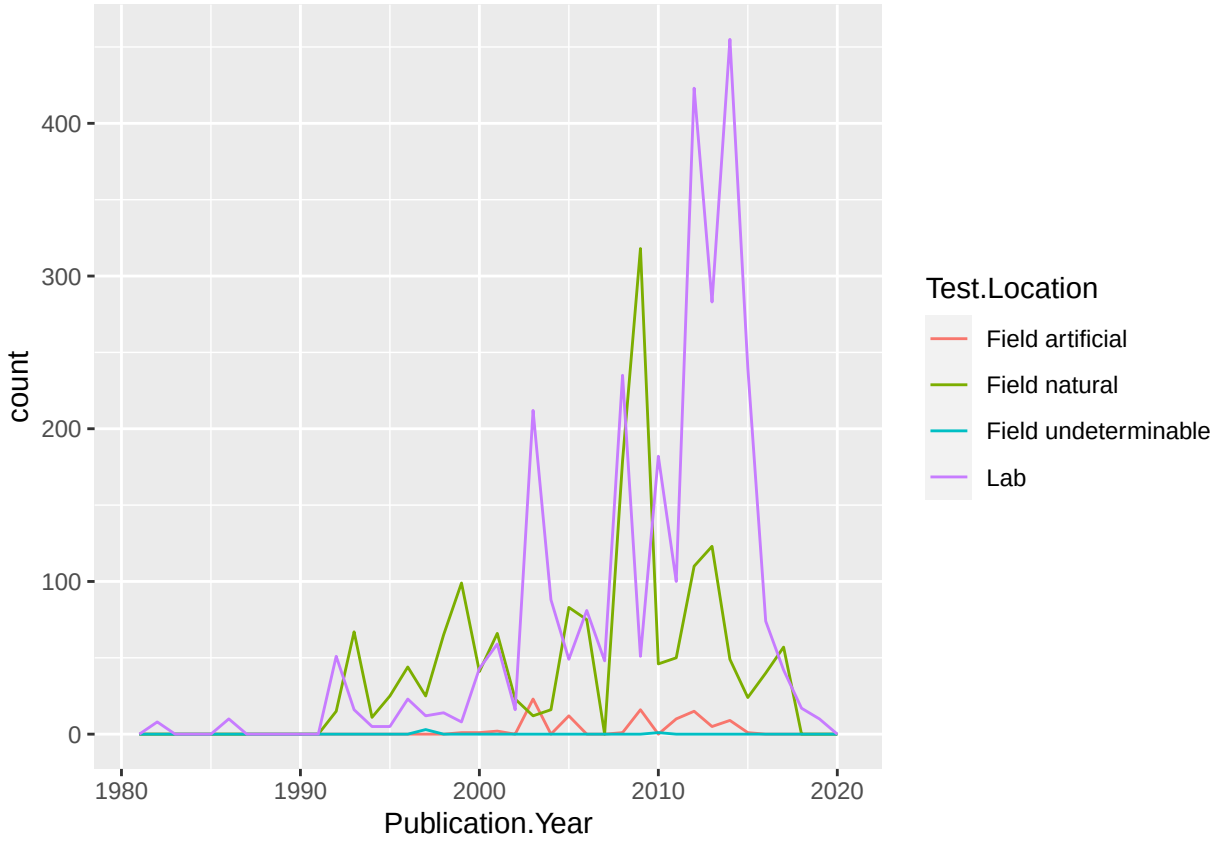
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics, aes(x = Publication.Year)) + geom_freqpoly(binwidth = 1)
```



10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
ggplot(Neonics, aes(x = Publication.Year, colour = Test.Location)) + geom_freqpoly(binwidth = 1)
```

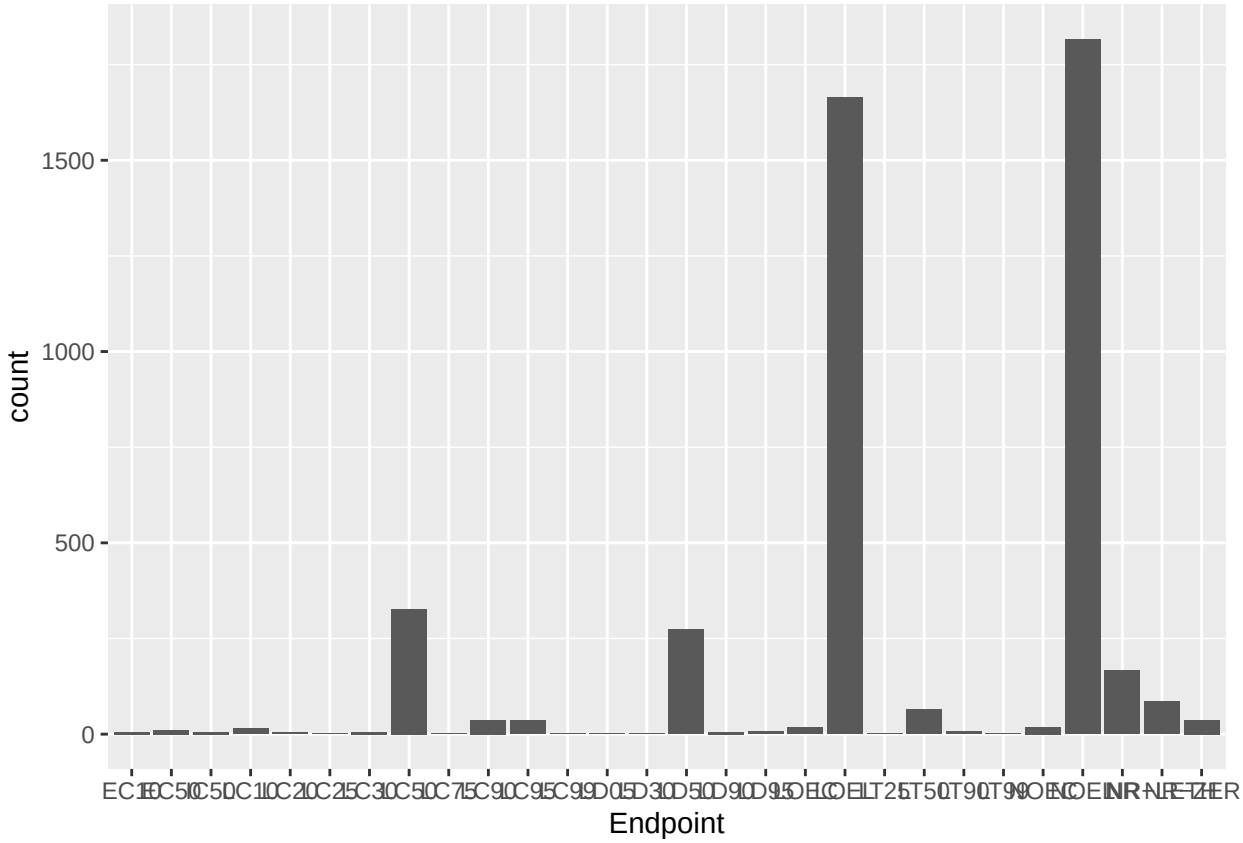


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer:The most common test locations are the Lab and natural in the field. All of them fluctuate around the same periods which probably pertains to times when greater priotity was placed on investigating the impact of pesticides on insects and there is also an overarching increase in lab experiements over time. This is likely due to improved practices on modeling realistic conditions within the lab.

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.

```
ggplot(Neonics, aes(x = Endpoint)) + geom_bar()
```



Answer: The two most common end points are LOEL and NOEL. LOEL is defined as Lowest-observable-effect-level or lowest dose producing effects that were significantly different. NOEL is defined as No-observable-effect-level or highest dose producing effects not significantly different from controls.

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"
```

```
# The class was not a date - it was a factor.

Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")

class(Litter$collectDate)
```

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

```
# The class was successfully changed to Date.

library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
Litter <- mutate(Litter, collectMonth = month(collectDate))
Litter <- mutate(Litter, collectYear = year(collectDate))

print(Litter$collectMonth == 8)
```

```
##   [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [46] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [61] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [76] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [91] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [106] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [121] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [136] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [151] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [166] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [181] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
print(Litter$collectYear == 2018)
```

```
##   [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [46] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [61] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [76] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [91] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [106] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [121] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [136] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [151] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [166] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [181] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
# since all outputs for the year being 2018 and the month being August was
# true, I have confirmation that the dataset is made up of data from August
# 2018 exclusively. I now only need to look at the unique dates for the dataset
# to understand the specific dates.

print(unique(Litter$collectDate))
```

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

```
# Litter was sampled on August 2nd and 30th in 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`?

```
print(Litter$siteID == "NIWO")
```

```
##   [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
##  [46] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [61] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [76] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [91] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [106] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [121] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [136] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [151] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [166] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [181] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```
print(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
```

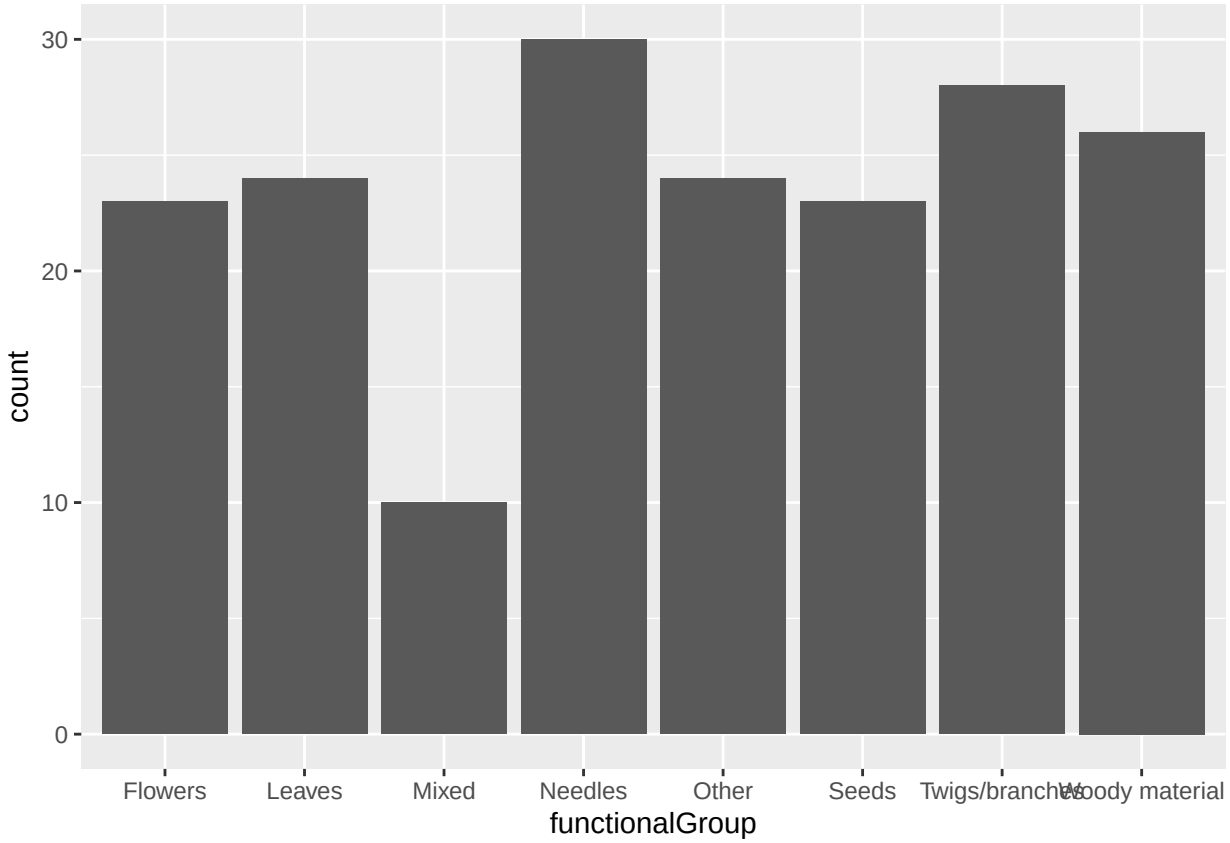
```
# 12 different plots were sampled at Niwot Ridge.

print(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 difference between unique and summary as functions is that unique merely lists out all of the different locations and summary tells you the number of occurances of each location.

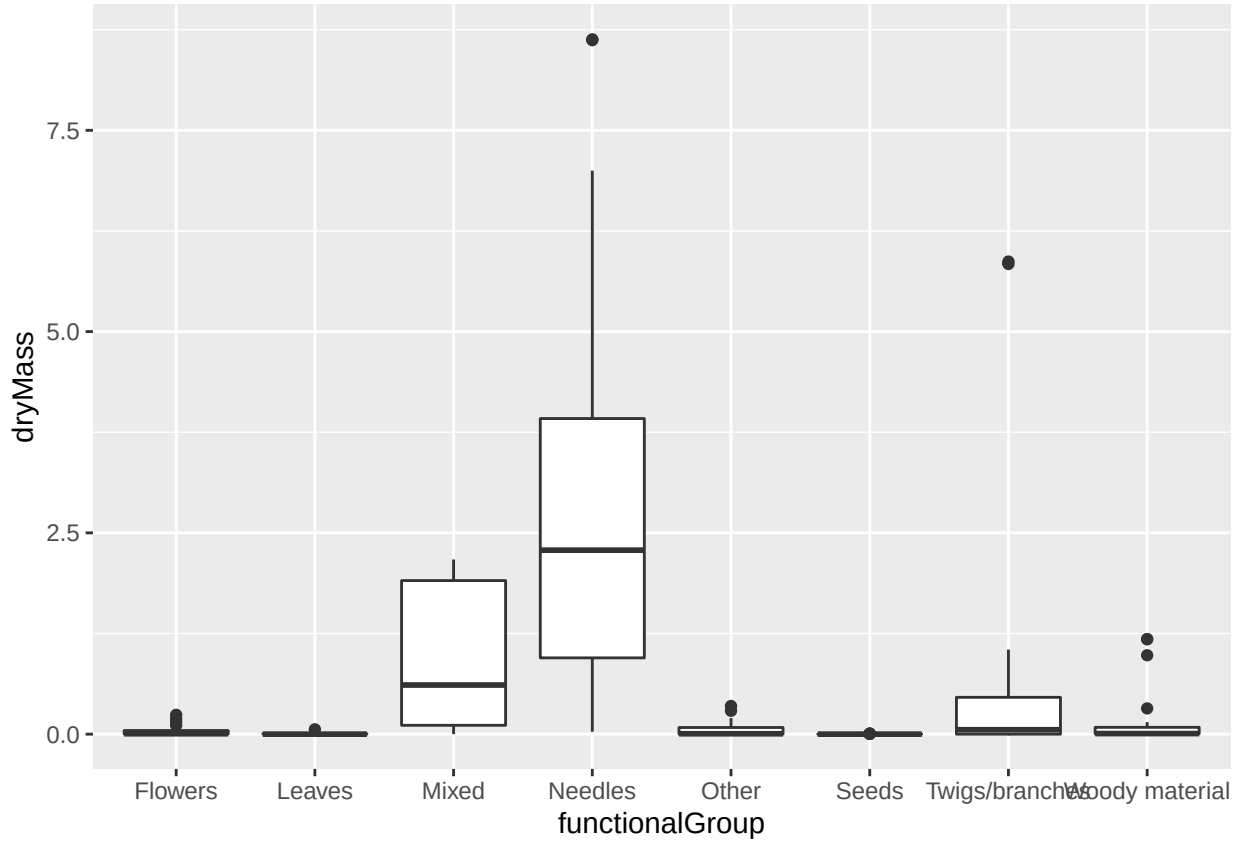
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

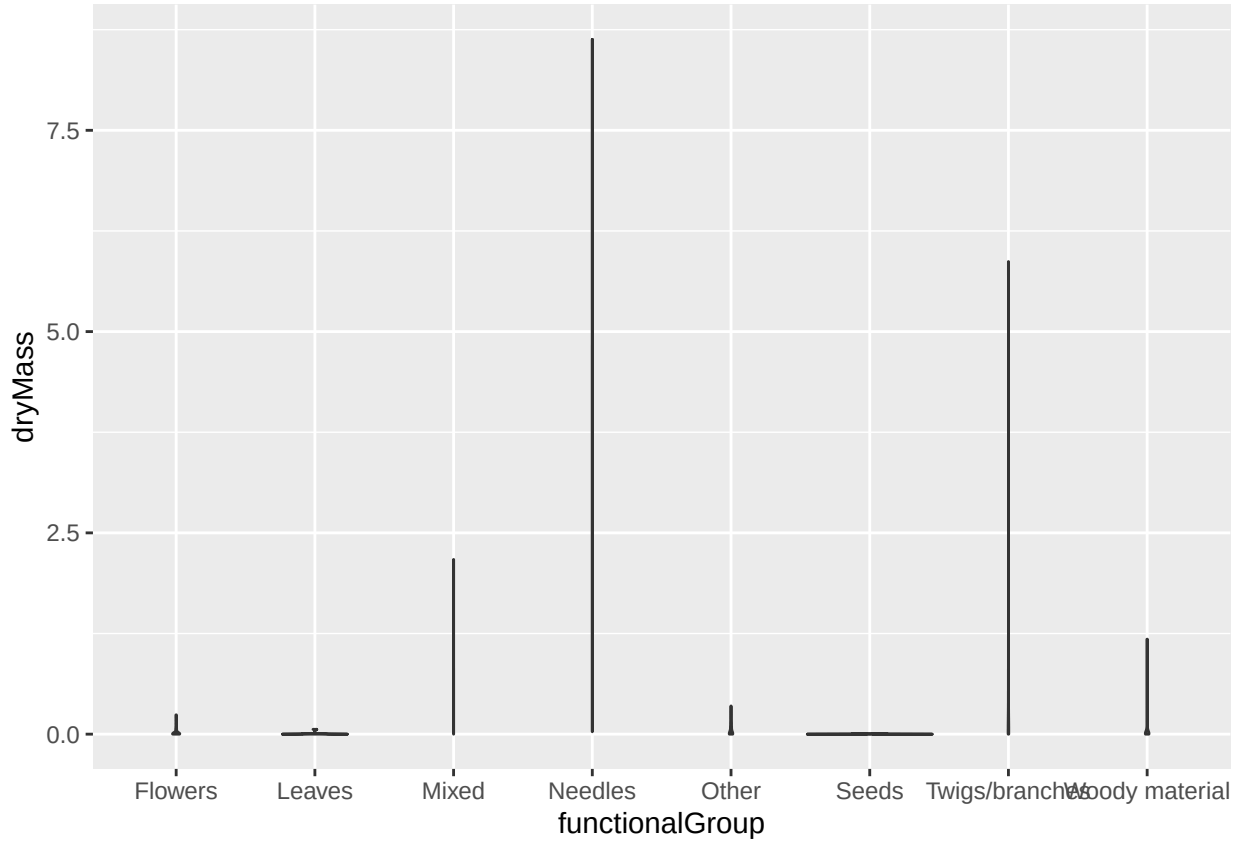


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




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



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

> Answer:The drymass value is either extremely densely overlapping or extremely spread out depending on the material. There are also not very many data total data points for each functional group. A boxplot marks and differentiates average regardless of how many points occupy the area. As the data is not robust enough for patterns to be highlighted by density of datapoints, a boxplot is necessary because it can still highlight general trends and designates outliers.

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

> Answer: Needles