Assignment 3: Data Exploration

Taylor Coleman

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.

```
#Loading tidyverse and lubridate
library(tidyverse)
library(lubridate)

#Uploading ECOTOX neonicotinoids dataset and Niwot Ridge NEON dataset
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: We might be interested in the ecotoxicology of neonicotinoids on insects because, similar to our concern about human health effects related to the use of insecticides for treating crops, insecticides have the potential to affect other insects aside from the target "pest(s)." For example, the use of these insecticides could have undesired negative impacts on local pollinators that are depended upon to aid in the reproduction of the crops being grown!

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: We might be interested in studying litter and woody debris that falls to the ground in forests because this litter/debris buildup can become hazardous fuel in the event of a forest fire. Conversely, data on litter and woody debris can be used as indicators for aboveground net primary productivity and biomass levels for the forest ecosystem being studied.

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: Litter and woody debris are sampled using both elevated and ground traps and then categorized into one of eight predefined functional groups, where 1. Litter is defined as material that has fallen from the forest canopy and has a butt end diamater less than 2 cm and length less than 50 cm that has been collected in 0.5mx0.5m elevated PVC traps. 2. Woody debris is defined as material that has fallen from the forest canopy and has a butt end diameter less than 2 cm and length greater than 50 cm that has been collected in ground traps. 3. Sampling occurs at terrestrial NEON sites that feature woody vegetation that is over 2 meters in height with trap pairs (one elevated, one ground) being placed either randomly or in targeted areas depending on vegetation coverage within 400mx400m plot areas.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#Having R return the number of rows and the number of columns for the Neonics dataset.
#The dimensions of this dataset are 4,632 rows x 30 columns.
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?

```r
#Extracting data from the "Effect" column of the Neonics dataset
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 appear to include population, mortality, and feeding behavior, ranked by descending order of their number of observations. These effects might be of particular interest because impacts they are all interrelated: detrimental impacts from the use of neonicotinoids on crops that these insects pollinate/consume will have a ripple effect in killing off individuals which in turn affects the entire population and increases mortality rates and, over time, will eventually result in a change in feeding behavior (for example, a bee colony is forced to identify a new food source and relocate their hive away from the cropland).

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...]

```r
#Use of "summary" function to determine 6 most commonly studied insect species from the Neonics dataset
sort(summary(Neonics$Species.Common.Name), decreasing = TRUE)
```

```
##                          (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 insect species from the Neonics dataset, based on their common names, include the honey bee, parasitic wasp, buff tailed bumblebee, Carniolan honey bee, bumble bee, Italian honeybee, and Japanese beetle. However, it is worth noting that the "Other" category had the greatest number of observations (670 whereas the honey bee had the second-highest at 667)! All of these species are pollinators - I learned that beetles are also considered pollinators, being called "mess and soil pollinators" by the Xerces Society for Invertebrate Conservation based on their messy eating habits - collecting and transferring pollen between the plants that they feed on. Thus, these insects, in comparison to others within the dataset, are more vulnerable to the application of insecticides on crops since they are directly reliant on them as a primary food source.

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?

```
#Determining the class of "Conc.1..Author" column in the Neonics dataset
class(Neonics$Conc.1..Author.)
```

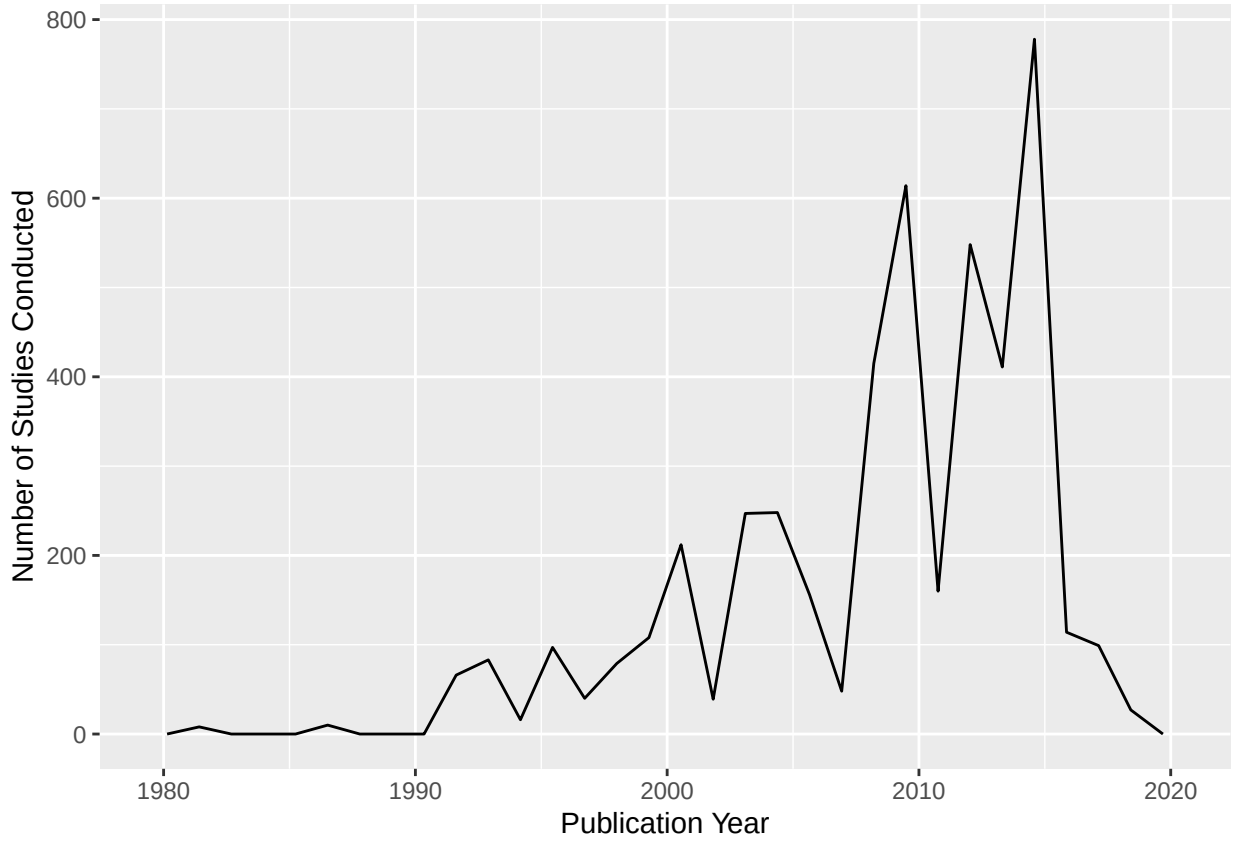
```
## [1] "factor"
```

Answer: Despite concentrations always being documented as a numeric value, R reports the class type for "Conc.1..Author" as a "Factor" because we specified with the "stringsAsFactors = TRUE" command when loading the Neonics dataset that we wanted R to read-in factors, which it automatically assigns to any vector that does not feature integer values.

Explore your data graphically (Neonics)

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

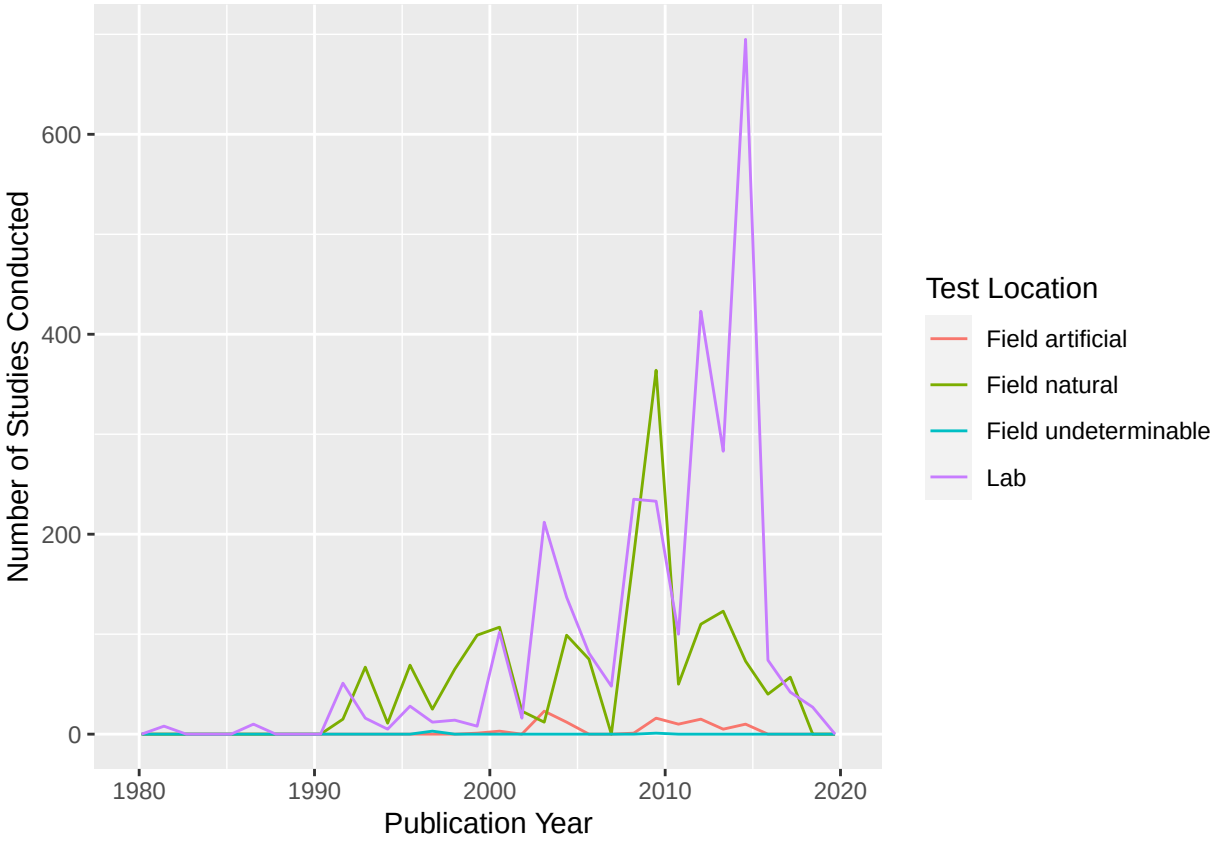
```
#Constructing a frequency line graph from Neonics to display studies conducted by publication year
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 30) +
  labs(x = "Publication Year", y = "Number of Studies Conducted")
```



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

```
#Reproducing the previous frequency line graph but adding display of test locations in different colors
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location)) +
  labs(x = "Publication Year", y = "Number of Studies Conducted") +
  labs(color ="Test Location")
```

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



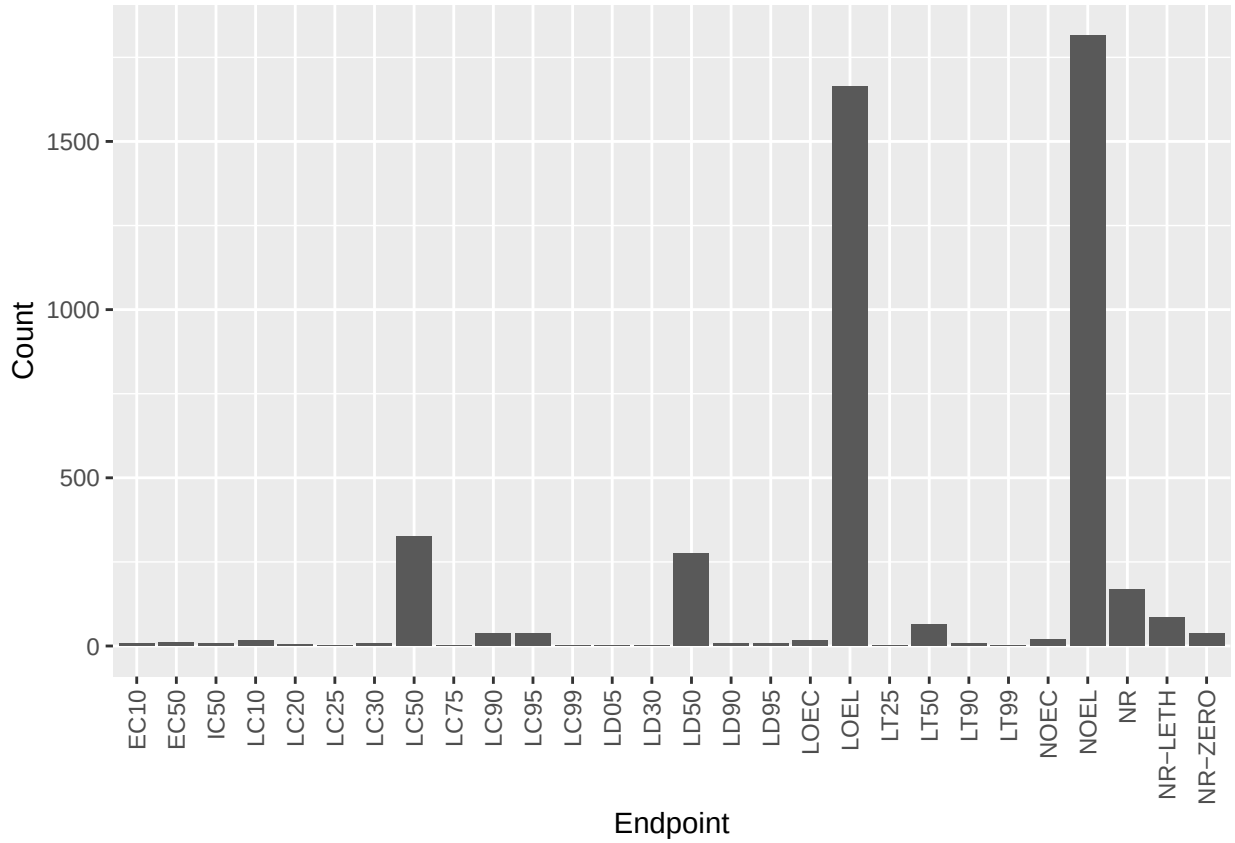
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are "lab" and "field natural" and yes, they vary over time. "Lab" begins as the most common test location, then is overtaken by "field natural" between 1990 and 2000. After the early 2000s, "Lab" takes the lead again until 2010, when "field natural" reaches its peak number of studies (close to 400). However, after 2010, close to 2015, lab reaches its own peak, being used as the location setting for nearly 700 studies.

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...]

```
#Creation of a bar graph for Neonics Endpoint counts
ggplot(Neonics) +
  geom_bar(aes(x = Endpoint)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  labs(x = "Endpoint", y = "Count")
```




```
#Checking findings from my bar graph
summary(Neonics$Endpoint)
```

```
##     EC10     EC50     IC50     LC10     LC20     LC25     LC30     LC50     LC75     LC90
##        6       11        6       15        5        1        6      327        1       37
##     LC95     LC99     LD05     LD30     LD50     LD90     LD95     LOEC     LOEL     LT25
##       36        2        1        1      274        6        7       17     1664        1
##     LT50     LT90     LT99     NOEC     NOEL       NR  NR-LETH  NR-ZERO
##       65        7        2       19     1816      167       86       37
```

Answer: NOEL and LOEL are the two most common endpoints from neonicotinoid toxicity testing. NOEL is defined as the no-observable-effect-level, or the highest concentration that produced effects that were not significantly different from those under the control responses, whereas LOEL is the lowest-observable-effect-level, or the lowest dose that produced significantly different effects from control responses.

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.

```
#Determining the class of collectDate from the Litter dataset.
#It is not recognized by R as a date (it is a "factor")!
class(Litter$collectDate)
```

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

```
#Changing collectDate so that it can become a date class
collectDate <- ymd(Litter$collectDate)

#Confirming that collectDate is now recognized as a "date"!
class(collectDate)
```

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

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`?

```
#Use of unique function to inspect the sample plots at Niwot Ridge
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
```

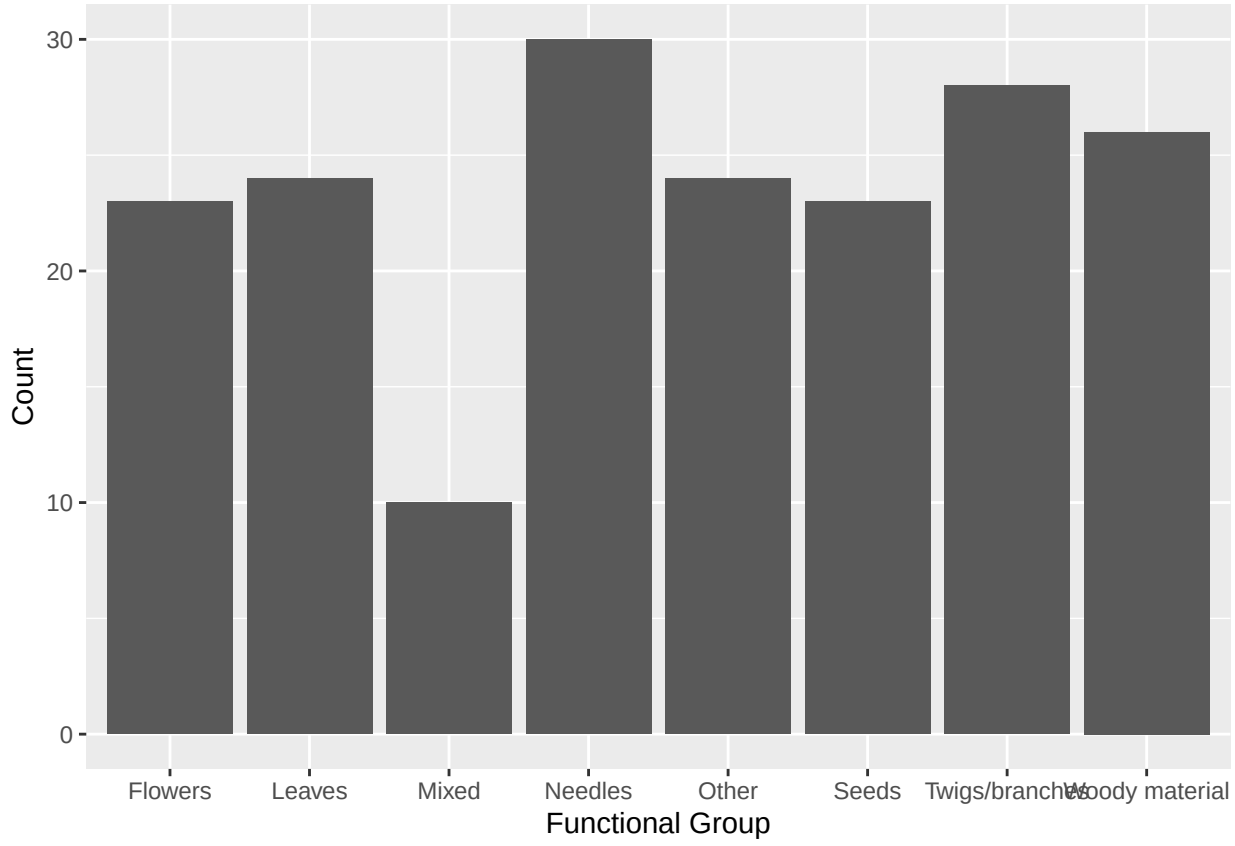
```
#Comparing the use of summary to inspect plot IDs
summary(Litter$plotID)
```

```
## 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: Whereas "summary" returns the number of observations associated with each plot ID, "unique" simply extracts unique elements from the Litter dataframe, returning all unique elements/rows for plot IDs by their name.

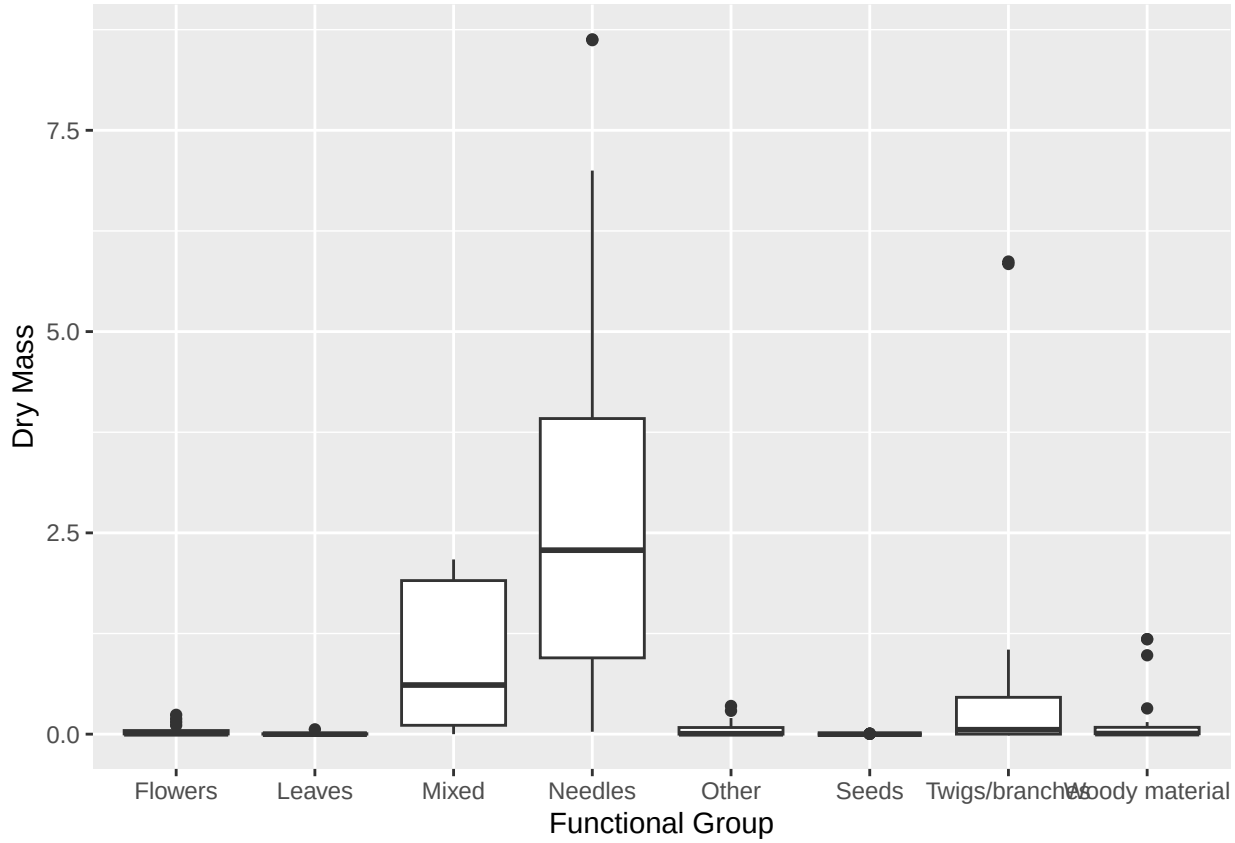
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#Constructing a bar graph to visualize the count per functional group from Litter
ggplot(Litter) +
  geom_bar(aes(x = functionalGroup)) +
  labs(x = "Functional Group", y = "Count")
```

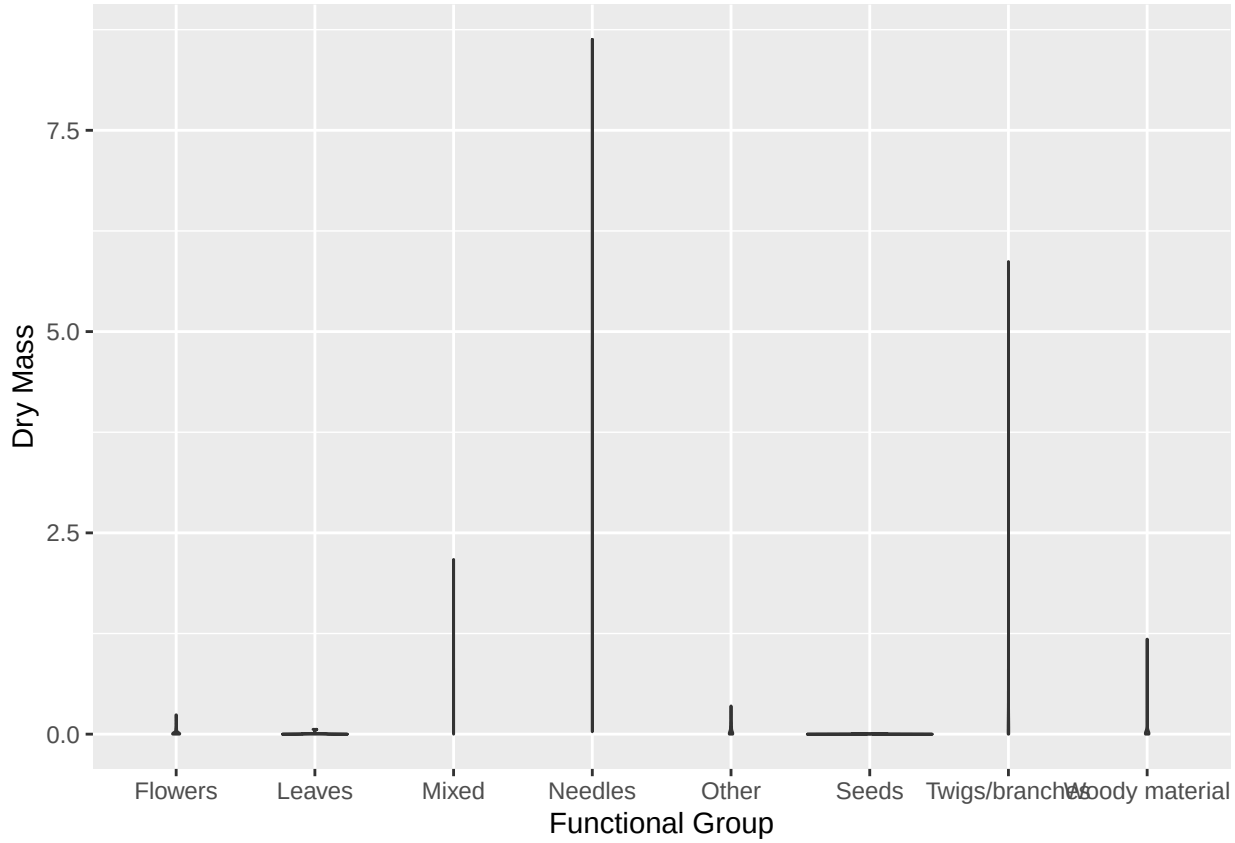


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#Constructing a boxplot to see the range of Litter dry mass values as they pertain to functional groups
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass)) +
  labs(x = "Functional Group", y = "Dry Mass")
```




```
#Constructing a violin plot for comparison of the same values
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass)) +
  labs(x = "Functional Group", y = "Dry Mass")
```



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

> Answer: The boxplot is a more effective visualization option than the violin plot in this case because it allows us to effectively see the interquartile range and median values for "mixed" and "needles" functional groups (the two groups featuring the largest IQR and median values) whereas the violin plot is able to show the range of values for these two functional groups but not their median values or the density distribution of each functional group's values. The violin plot is including outliers within its depicted ranges, drawing attention to "twigs/branches" as the functional group with the second largest range of values to needles, leaving no knowledge about the narrow IQR that it has relative to the "mixed" functional group, as revealed by the box plot.

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

> Answer: Needles and mixed litter tend to have the highest dry biomass at these sites based on the reported median dry mass value depicted on the box plot.