Assignment 3: Data Exploration

Sarah Sussman

Spring 2025

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 Canvas.

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. Load necessary packages (tidyverse, lubridate, here), check your current working directory 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.

```
# Import packages
library(tidyverse)
library(lubridate)
library(here)

# Check current working directory
getwd()
```

```
## [1] "/home/guest/EDA_Spring2025"
```

```r
# Upload necessary datasets
Neonics <- read.csv(here('Data', 'Raw', 'ECOTOX_Neonicotinoids_Insects_raw.csv'),
                    stringsAsFactors = T)
Litter <- read.csv(here('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: We know that neonicotinoids are insecticides, thus they target insects. However, certain neonicotinoids may be harmful to non-target insects, or insects that we actually don't mean to harm; I am specfically thinking of non-harmful pollinators like some species of beetles, butterflies, ants, etc. This can lead to "unintended consequences". An unregulated approach to using neonticotinoids could lead to a dwindling population of pollinator insect species. Studies on the effects of neonictinoids could inform how to regulate the insecticide and how to better formulate it to target the exact species of interest.

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: There are so many reasons! Some reasons that come to mind are to study species that live in forest floor, like: insect larvae, amphibious species (salamanders!), and spring ephemerals. Studying woody debris could also indicate forest health and answer specific research questions - for example how many cones is a stand of eastern hemlocks dropping? Is it still healthy after being treated for hemlock woolly adelgid? (at least in the southeast, not Colorado).

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. Litter debris is collected through elevated traps, while woody debris is collected through ground traps. 2. Depending on spatial arrangement of the forest, trap placement may be targeted or randomized. In places with

> 50% aerial cover with vegetation > 2m, trap placement will be randomized. In places with less aerial c
and shorter vegetation height, trap placement will be targeted to be under qualifying vegetation.

3. Ground traps are sampled once a year. Elevated trap sampling frequency depends on forest type and time of year, for example: in decidious forest types, elevated traps are sampled once every two weeks during senescence (autumn).

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
# Find dimensions of the dataset
dim(Neonics) # 4623, 30
```

```
## [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? [Tip: The `sort()` command is useful for listing the values in order of magnitude...]

```
# Determine the most common effects that are studied
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
```

```
#sort(Neonics$Effect, decreasing = TRUE)
```

Answer: Population and Reproduction are the two most common effects that are studied. These are probably the most studied because studying how neonicotinoids affects insect reproduction and population would be paramount to knowing how neonicotinoids affect species survival.

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: Explore the help on the `summary()` function, in particular the `maxsum` argument...]

```
# Determine the six most commonly studied species in the dataset (common name)
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 commonly studied species in the dataset are: Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. What they all have in common is that they are are bee species. They are of more interest compared to other species because they are pollinators of our agricultural systems. I assume that the agricultural sector is the biggest user of neonicotinoids to protect crops from insect pests. Because neonticotinoids target ALL insects, it would also harm benefical pollinator insect species, like bees.

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? [Tip: Viewing the dataframe may be helpful...]

```
# What class is 'Conc.1..Author'
class(Neonics$Conc.1..Author.)
```

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

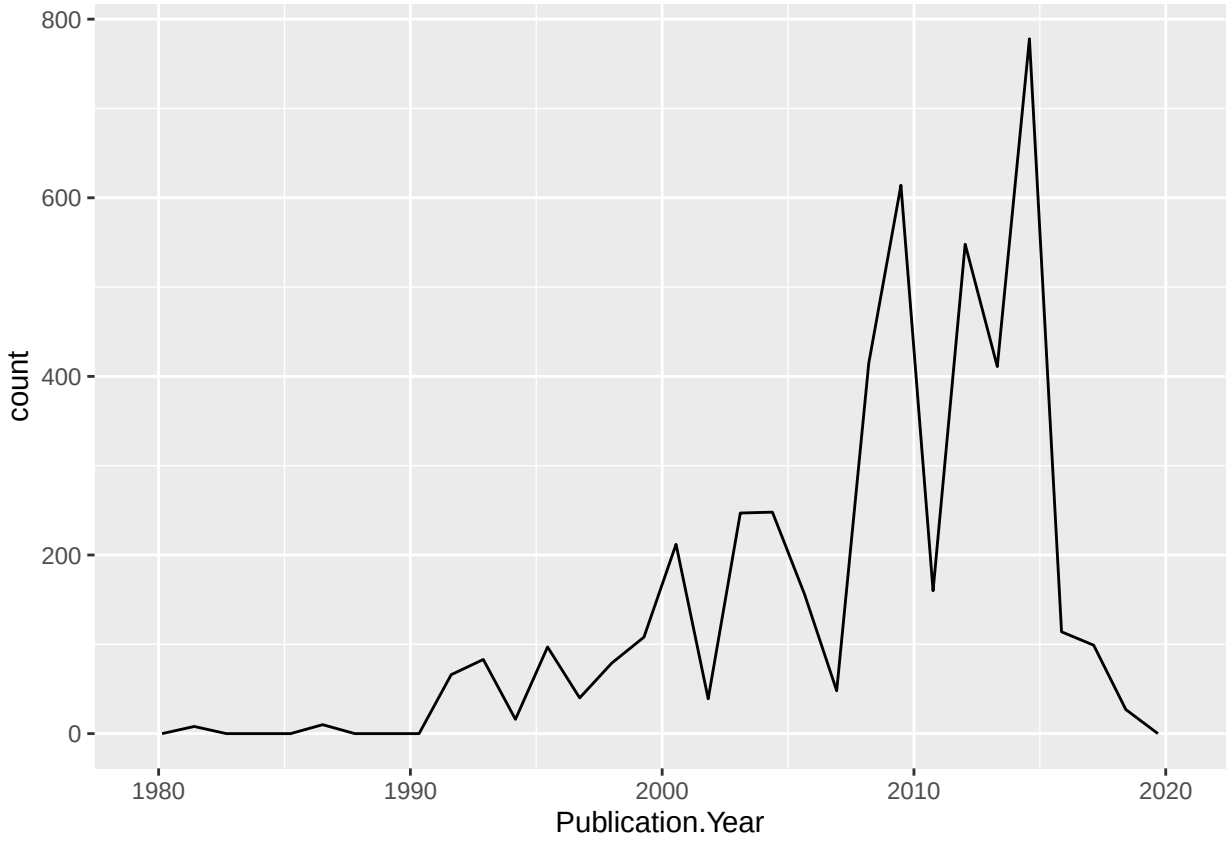
Answer: The class of 'Conc.1..Author' is factor. This column is not numeric because it is reporting a measurement, so the data is being stored as a factor.

Explore your data graphically (Neonics)

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

```
# Generate frequency line graph of the number of studies conducted by publication year
ggplot(Neonics) +
  geom_freqpoly(
    aes(x = Publication.Year))
```

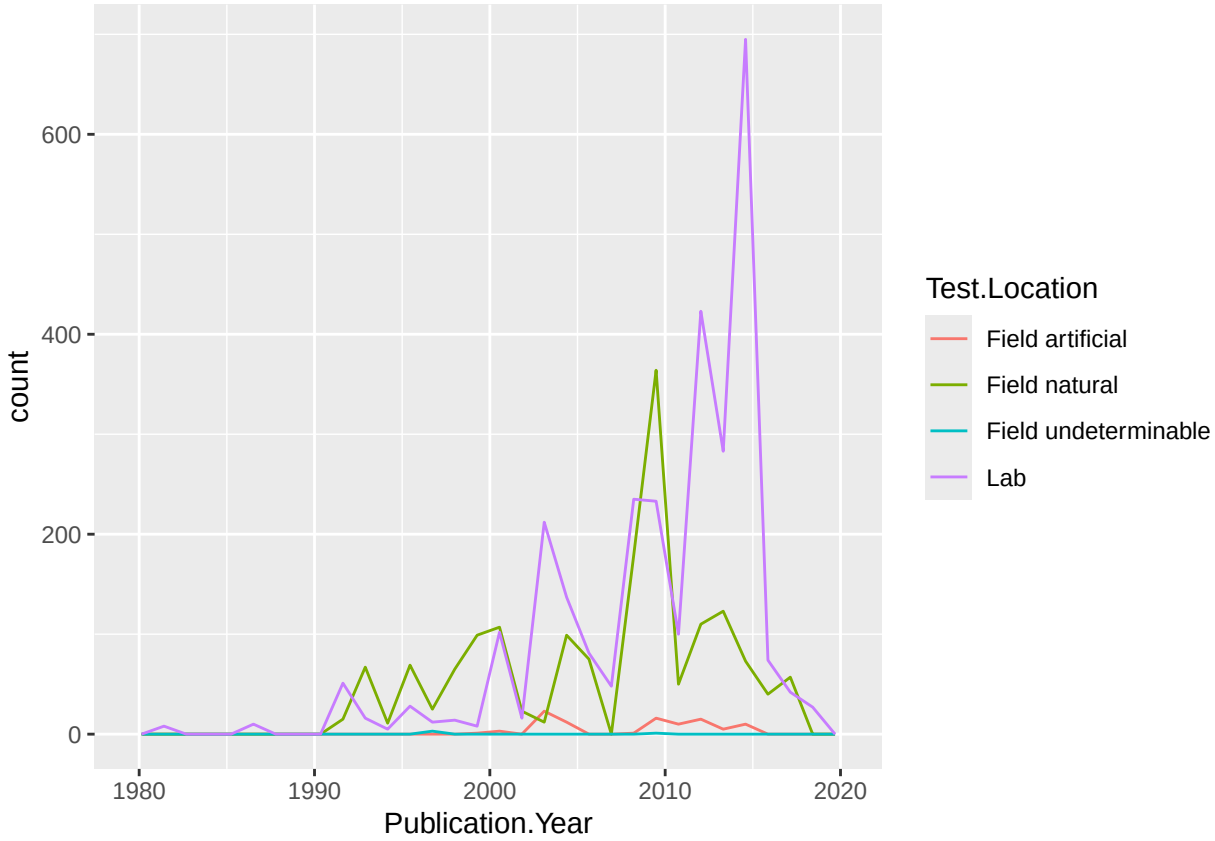
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
# Generate the same frequency line plot as above but color it by Test.Location
ggplot(Neonics) +
  geom_freqpoly(
    aes(x = Publication.Year, color = Test.Location))
```

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



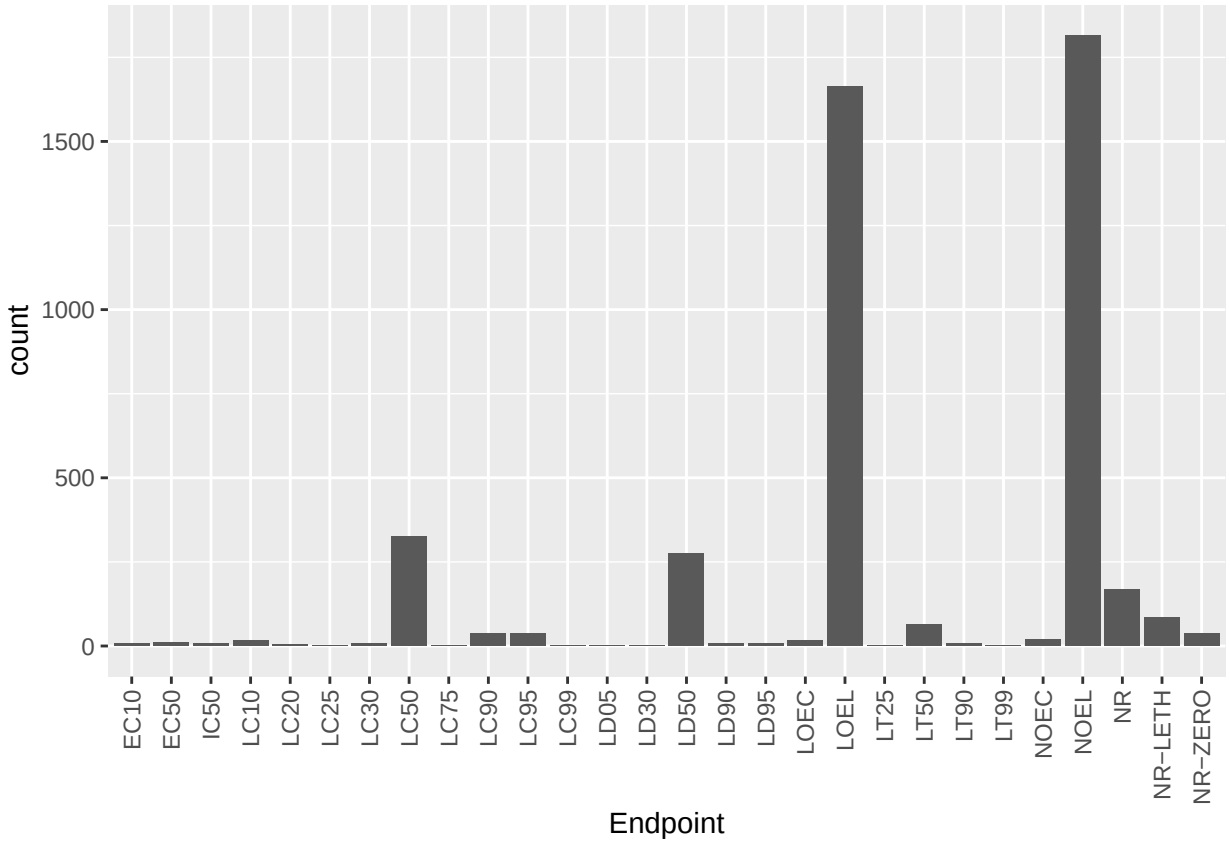
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common tesing locations are in the lab and in the field natural. The frequency of both has changed over time; in 2010 the field natural was the most common testing location and in 2015 the lab was the most common. The other two testing locations, field undeterminable and field artifical are not common at all - it seems like field undeterminable has never been a testing location (probably because it is clear where testing is done, and this testing location type is for when tesing location is not clear, which seems like it would be a rare occurence.)

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...]

```
# Generate a bar graph of the Endpoint counts
ggplot(Neonics,
       aes(x = Endpoint)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: The two most common end points are NOEL and LOEL. NOEL stands for "No-observable-effect-level: highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test". LOEL stands for "Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of 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) # collectDate is a factor
```

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

```
# Change to a date
Litter$collectDate <- ymd(Litter$collectDate)
class(Litter$collectDate) # collectDate is a dat
```

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

```r
# Using the 'unique' function, determine which dates litter was sampled in August 2018.
unique(Litter$collectDate) # 08/02/2018 & 08/30/2018
```

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

13. Using the `unique` function, determine how many different plots were sampled at Niwot Ridge. How is the information obtained from `unique` different from that obtained from `summary`?

```r
# Use the unique function to determine how many different plots were sampled 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
```

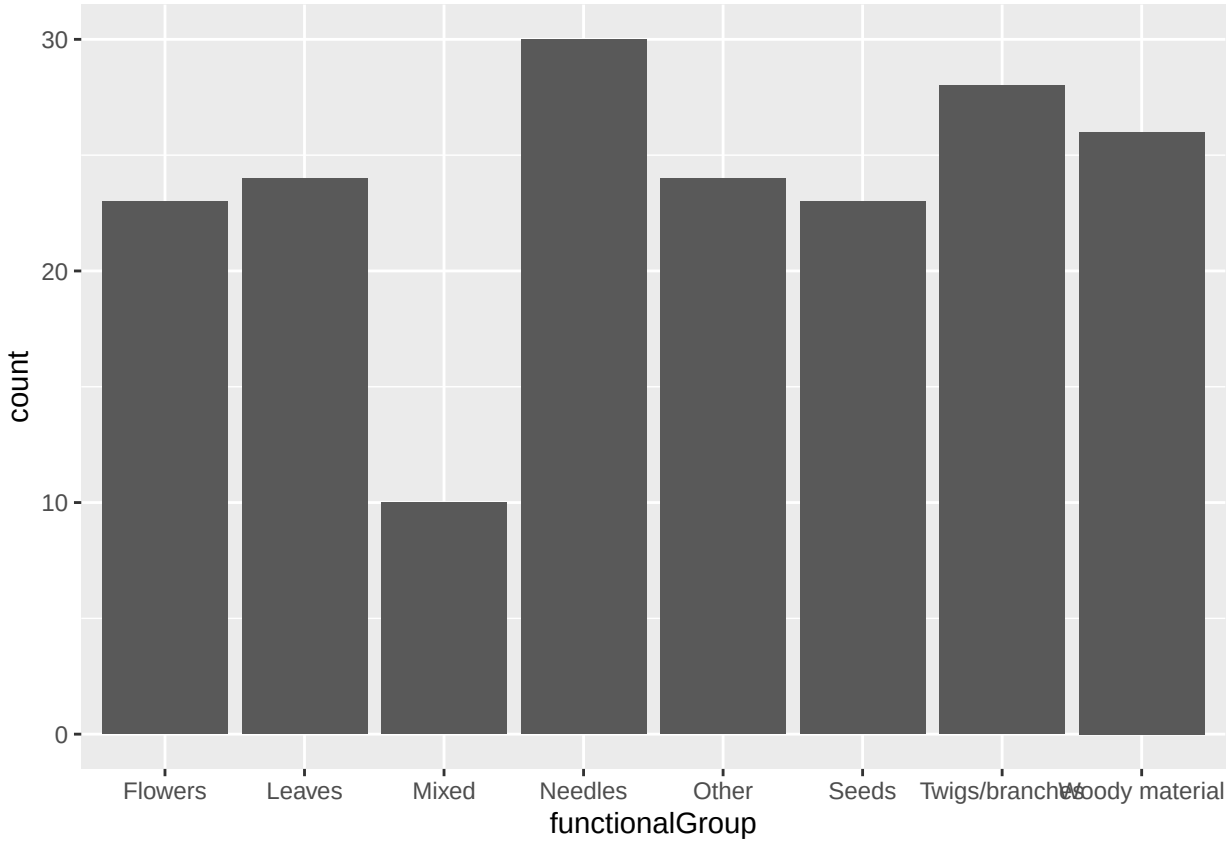
```r
summary(Litter$plotID)
```

```
## 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 unique function gives the name of each unique plot ID while the summary function gives the name of each unique plot ID and the number of times that plot ID occurs in the dataset (how many times it was sampled).

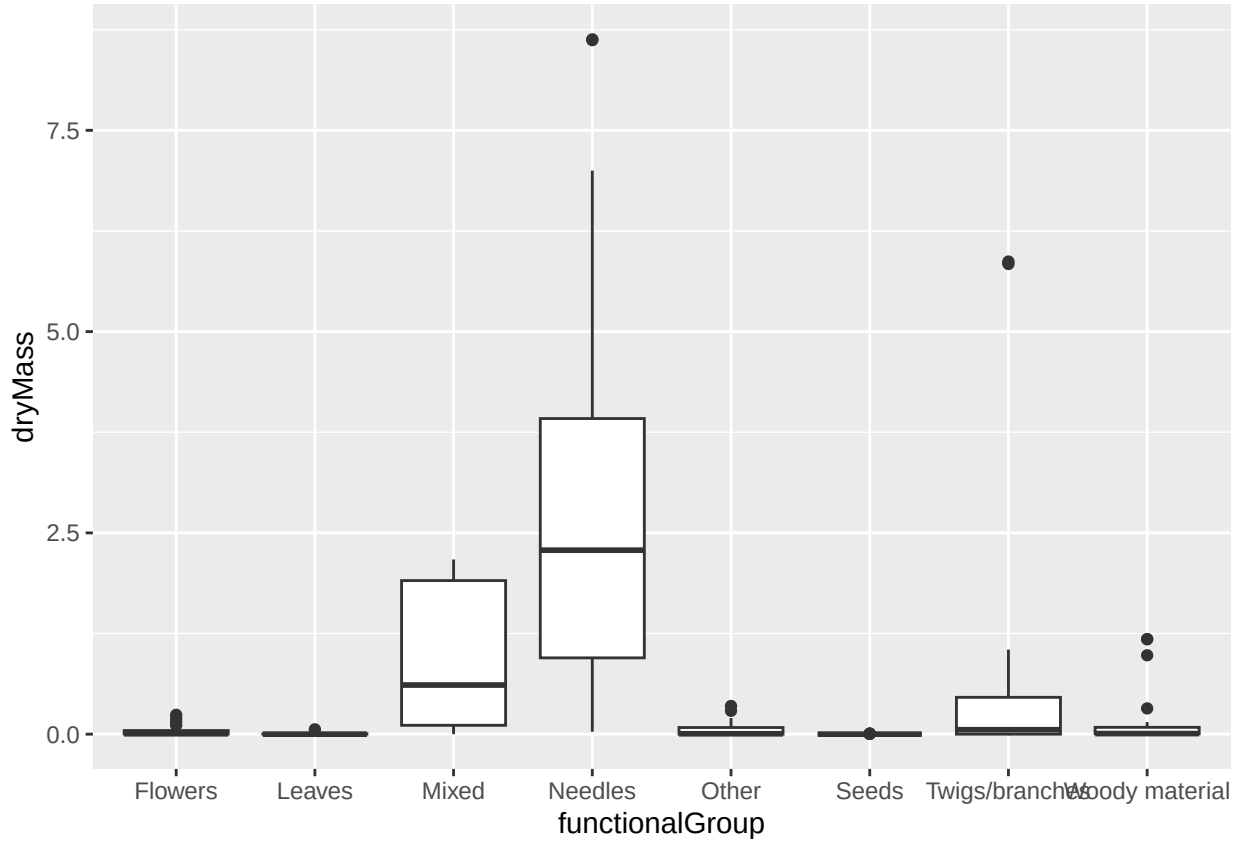
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
# Generate a bar graph of functionalGroup counts
ggplot(Litter,
       aes(x = functionalGroup)) +
  geom_bar()
```

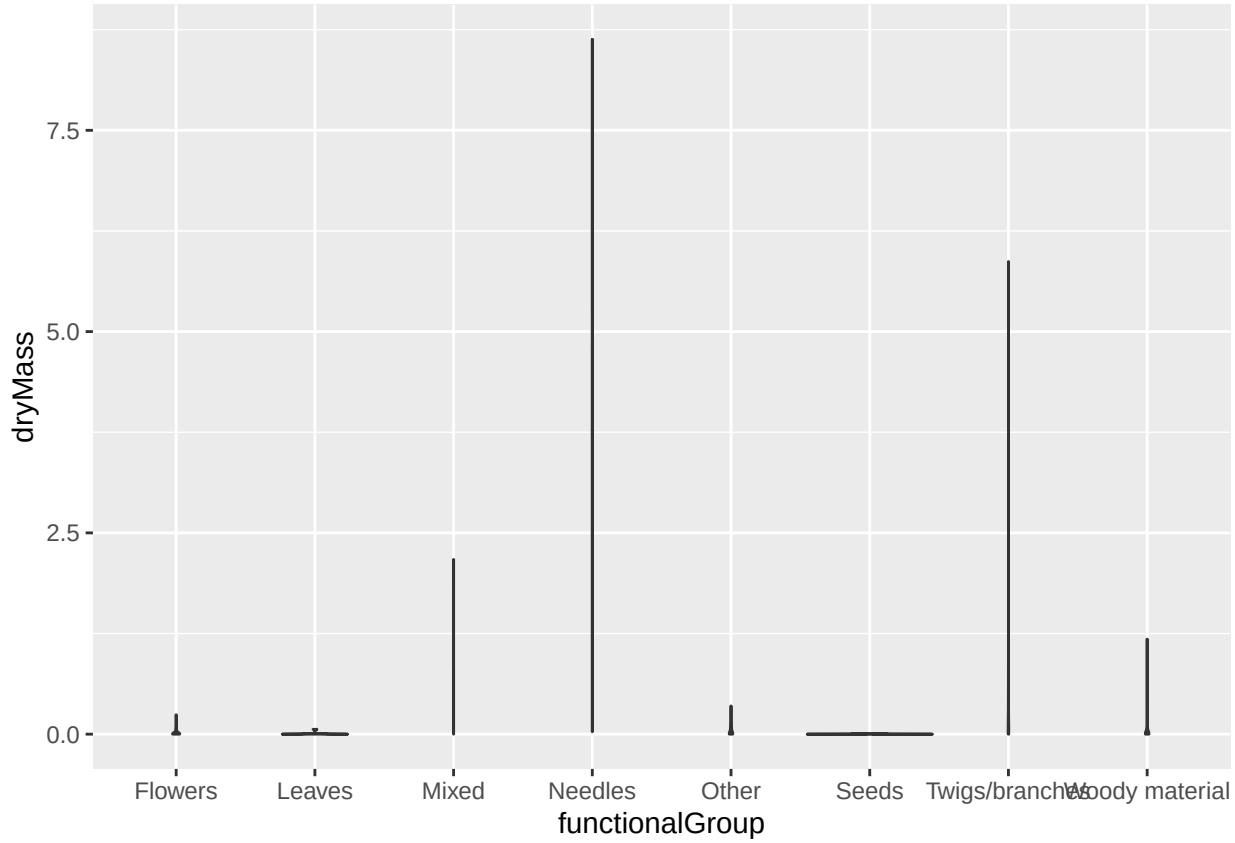


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
# Box plot of dryMass by functionalGroup
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup,
                   y = dryMass))
```




```
# Violin plot of dryMass by functionalGroup
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup,
                  y = dryMass))
```



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

> Answer: The violin plot just looks like thick lines, whereas the box plot shows some boxes (though it is not perfect), with the IQR, median, and outliers being displayed in the graph. Violin plots show the width as propotional to the number of values, reflecting density. Since the graph is just showing dryMass by functionalGroup, there isn't really any density to show.

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

> Answer: Needles tend have the highest biomass at the sites.