Assignment 3: Data Exploration

Angela Edwards

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)

# Import two datasets with read.csv function, adding stringsAsFactors command.
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: The effects of insecticides like neonicotinoids on insects, beyond the targeted insect, is a very important course of study, as researchers seek to discover unintended consequences of the chemicals. For instance, according to the New Jersey Audubon Society, neonicotinoids harm the immune systems of bees, weakening the ability of the bee colony to survive. Documented harm to insects may also be illustrative of harms that could travel up the food chain to humans.

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 study of litter and woody debris in forests is an important area of study. Researchers may be interested in a number of topics, including the following few examples: the matter's potential to start or spread forest fires; the rate of forest decay; the role of nutrient cycling and carbon budgets, the provision of habitat for varied species, and how nearby stream systems and aquatic life are affected by the cycling of the matter.

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 is collected from elevated traps (size: 0.5mˆ2 square with mesh basket elevated ~80cm above the ground), while fine woody debris is collected from ground traps (3m x 0.5m rectangular areas). 2. Sampling is done once per year for ground traps while sampling for elevated traps varies by vegetation and season, with decidiuos forest samples taken once every 2 weeks during senescence, and evergreen samples taken once every 1-2 months or less. 3. Traps are placed in pairs (one elevated and one ground trap) per every 400mˆ2 plot area. Trap placement within the plots may be targeted or randomized, depending on the vegetation.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
# Find the dimensions of the Neonics 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?

```
# Find the summary of the Effect column in 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 two most common effects have to do with Population and Mortality, which are important elements in understanding the mortal effects of the chemical on particular insect populations.

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. . .]

```
# Find summary of Species.Common.Name.
most_common_species <-summary(Neonics$Species.Common.Name)
sort(most_common_species, 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 species in the dataset are: Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. Five of the six species are bees, but all of the species could be considered crop pollinators, which may increase their interest to researchers over non-pollinating insects. Crop pollination is an important ecosystem service upon which many crops depend.

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?

```
# Find class of Conc.1..Author.
class(Neonics$Conc.1..Author)
```

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

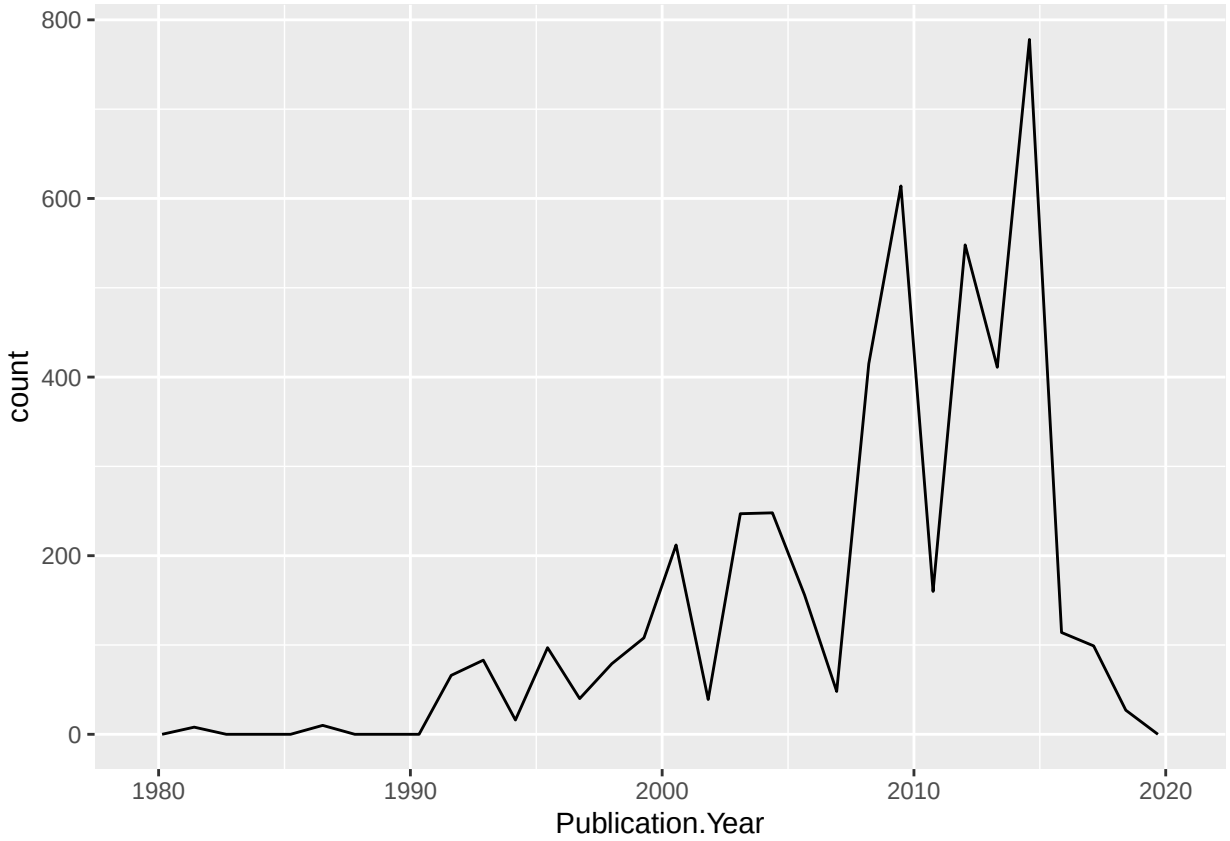
Answer: The class function tells us that the `Conc.1..Author` variable in the dataset is a factor. This variable is not numeric because (potentially?) in the context of the dataset, it should be considered categorical. Alternatively, since we imported the data specifying that strings should become factors, then the data was originally coded as a string.

Explore your data graphically (Neonics)

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

```
# Use geom_freqpoly to create a plot of the number of studies per publication year.
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year))
```

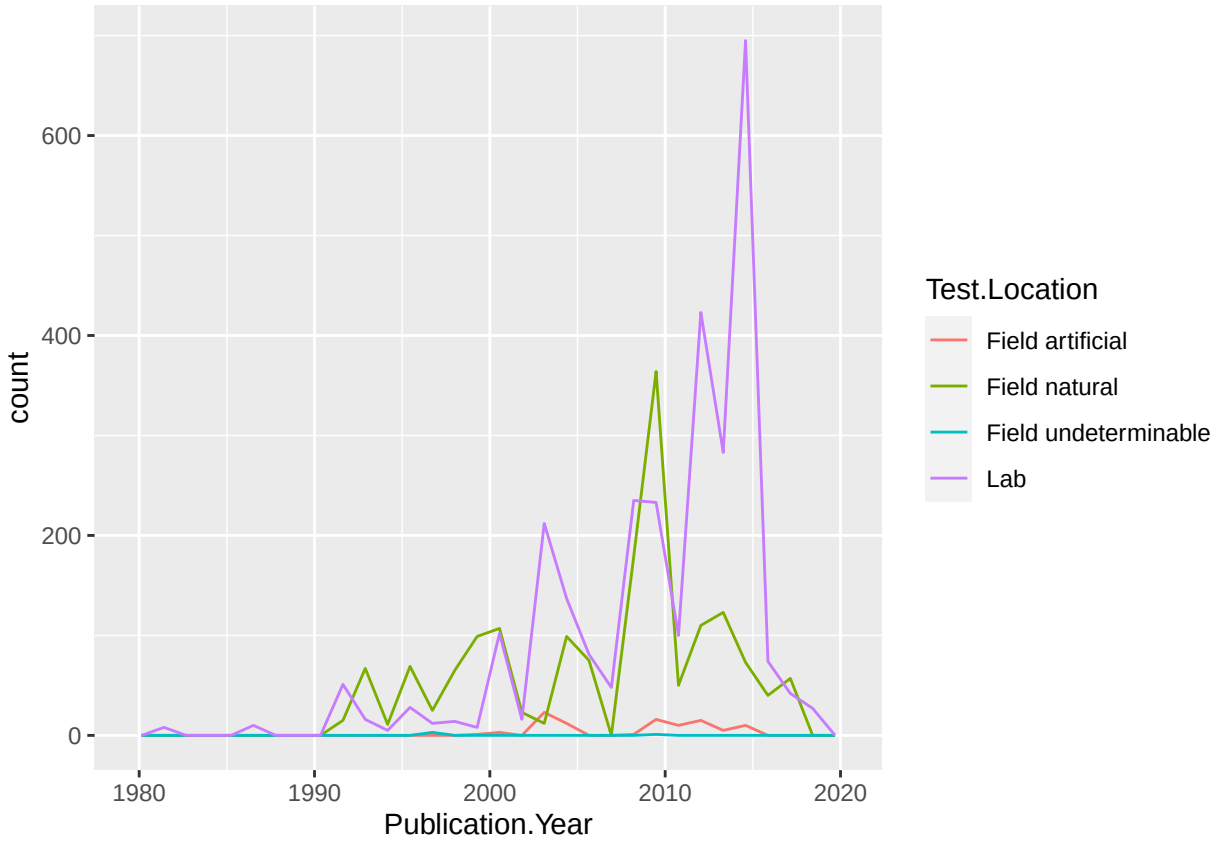
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
# Add Test.Location variable with color.
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location))
```

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



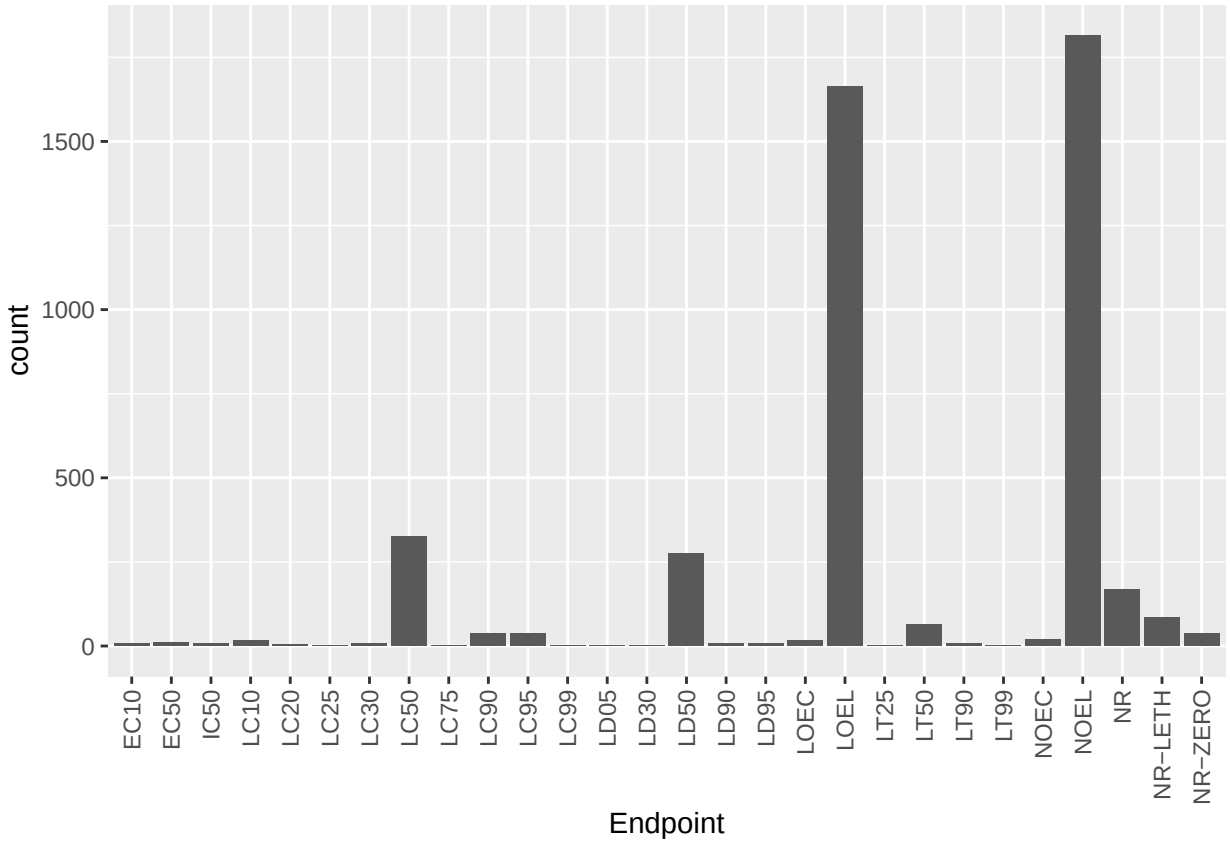
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The two most common test locations are `Lab` and `Field natural` and they vary in terms of use over time: the Lab location seems to be more prevalent from about 2003 to 2004, and then much more prevalent from about 2012 through 2016. The natural field location seems to be more common during about 1995 to 2001 and again around 2008/2009.

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...]

```
# Find the two most common end points.
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 endpoints are LOEL and NOEL. LOEL is defined as the lowest dose producting effects that were significantly different from control responses. NOEL is defined as the highest dose producting effects not significantly different 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.

```
# Find class of collectDate in Litter dataset.
class(Litter$collectDate)
```

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

```
# Change collectDate to datetime format.
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
class(Litter$collectDate)
```

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

```
# Find which dates litter was sampled in August 2018. (2018-08-02 and 2018-08-30)
unique(Litter$collectDate)
```

```
## [1] "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`?

```
# Use unique to find how many plots were sampled at Niwot Ridge.
# Contrast this result with the result of the summary function.
```

```
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
```

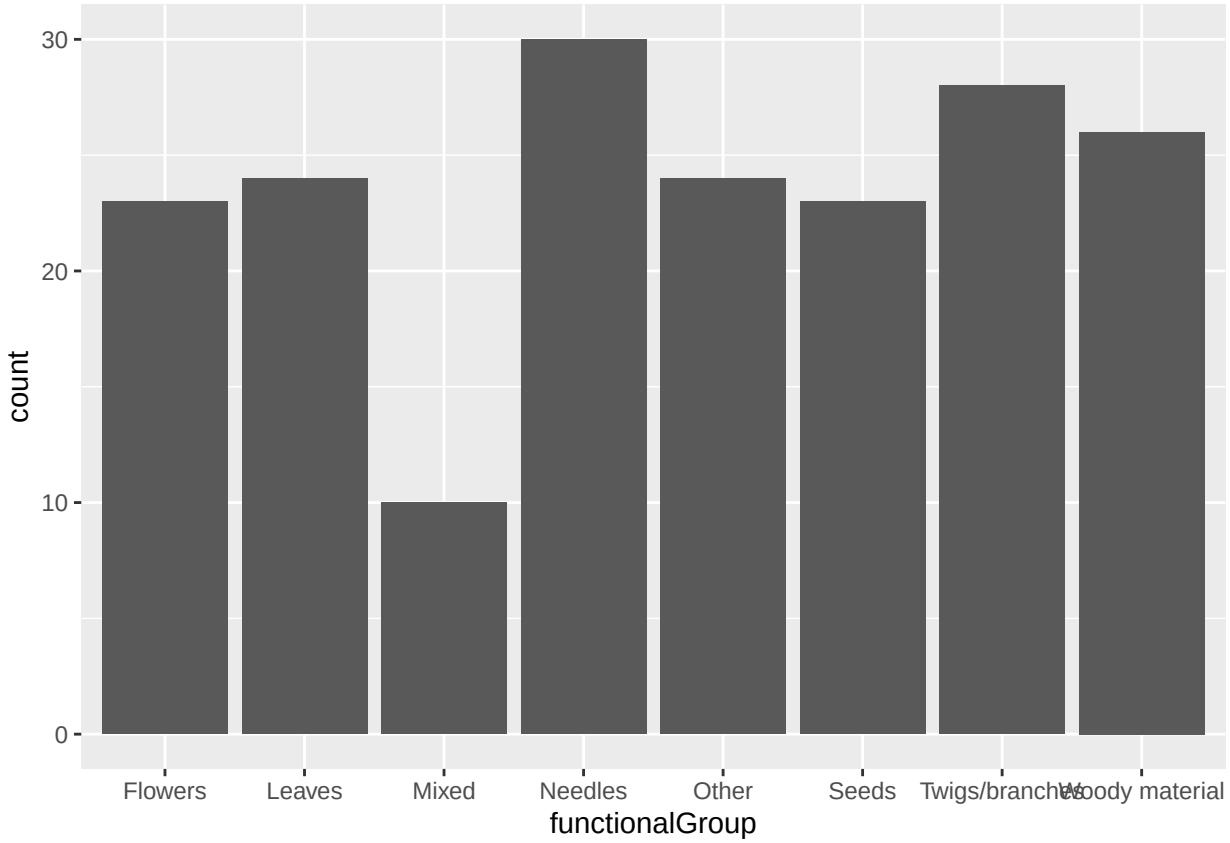
```
summary(Litter$namedLocation)
```

```
## 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 plots sampled at Niwot Ridge. The information obtained from the unique function lists the unique values of the locations. Using the summary function, we can see each of the 12 values, but we also see how many observations fall within each of the 12 plots.

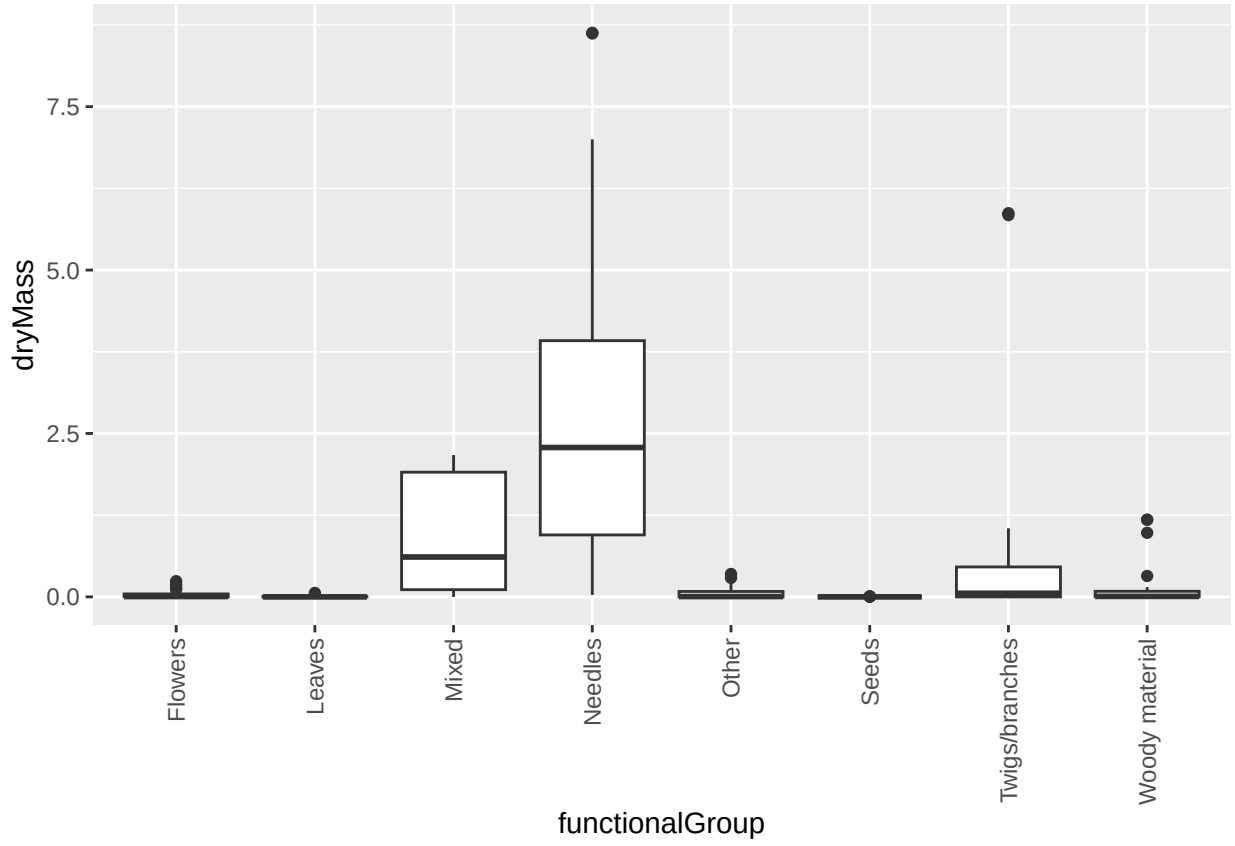
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
# Create a bar graph of functionalGroup counts from Litter dataset.
ggplot(Litter) +
  geom_bar(aes(x=functionalGroup))
```

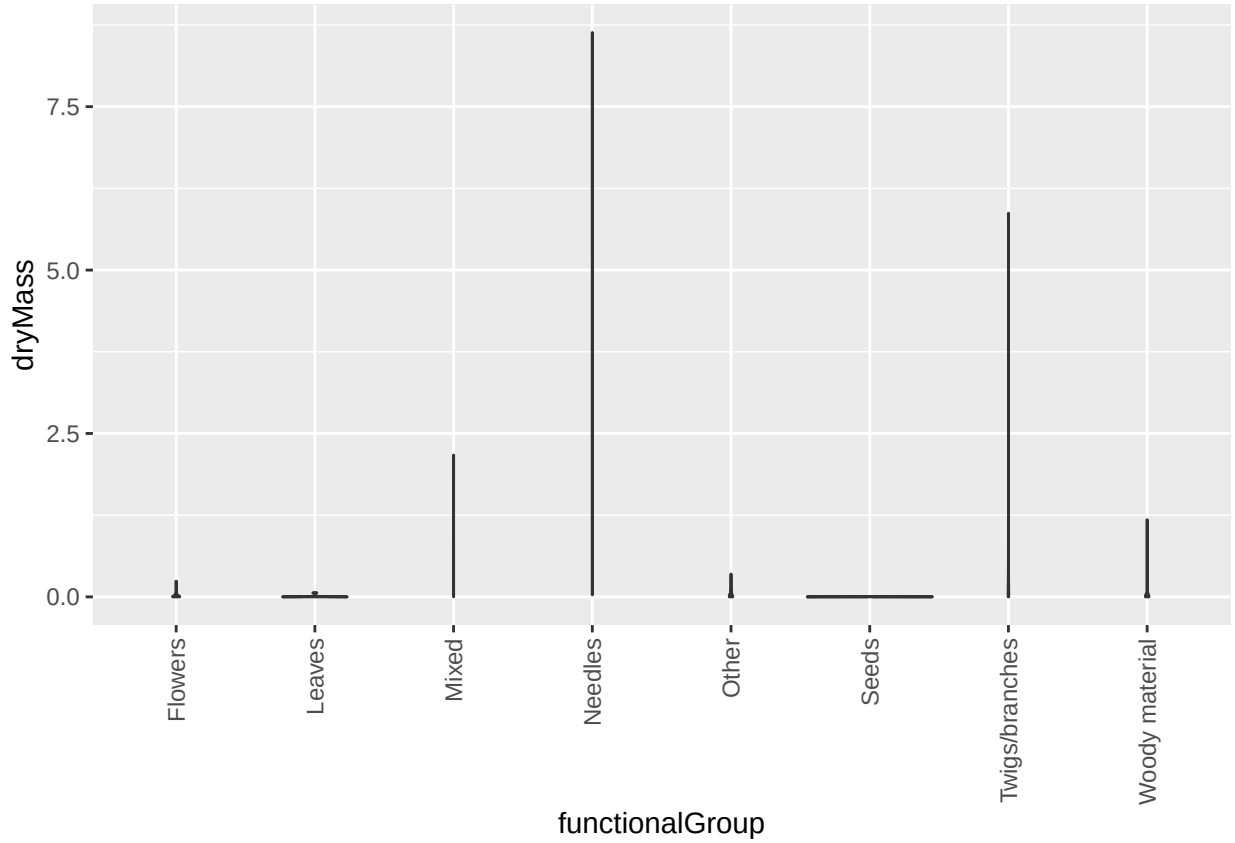


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
# Create boxplot of dryMass by functionalGroup.
ggplot(Litter) +
  geom_boxplot(aes(x=functionalGroup, y=dryMass)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```




```
# Create violin plot of dryMass by functionalGroup.
ggplot(Litter) +
  geom_violin(aes(x=functionalGroup, y=dryMass)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



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

> Answer: Using a violin plot would normally add in a little extra distribution, but since the litter types are fairly equally distributed, the violin plot is not as effective as 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 biomass at these sites.