Assignment 3: Data Exploration

Xiaoge Zhang

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.

```
# check work directory
getwd()
```

```
## [1] "/home/guest/R/EDA-Spring2023"
```

```
# load packages
library(tidyverse)
library(lubridate)

# upload datasets
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 intersted in finding out more about how neonicotinoids pass down on food webs and how the insects, particularly pollinator species, could be potentially impacted by the neonicotinoids.

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 learning about litter and woody debris that falls to the ground to to know more about ground covers and the well-being and productivity of ecosystems; these information could also help us estimate aboveground carbon/biomass and other important ecological indices such as the leaf area index.

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.Spatial scale: Samplings occurred in tower plots, and tower plots were selected randomly within certain areas. Trap placements were targeted or randomized depending on vegetation. 2.Temporal scale: Ground traps were sampled once a year; sampling frequency for elevated traps varied for vegetation and season. 3. Both spaital and temporal resolution and extent were clearly defined. The finest resolution of temporal data is days of trapping; the finest spatial resolution is a single trap.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
# dimentiobs of 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?

```
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 effect that is studied is population. The second most common one is mortality. These effects might be specifically of interest because insect population and mortality might affect the food webs and the well-being of our ecosystems drastically.

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…]

```
sort(summary(Neonics$Species.Common.Name))
```

```
##                          Ant Family                        Apple Maggot 
##                                   9                                   9 
##               Glasshouse Potato Wasp                            Lacewing 
##                                  10                                  10 
##             Southern House Mosquito             Two Spotted Lady Beetle 
##                                  10                                  10 
##             Spotless Ladybird Beetle                 Braconid Parasitoid 
##                                  11                                  12 
##                        Common Thrip        Eastern Subterranean Termite 
##                                  12                                  12 
##                              Jassid                          Mite Order 
##                                  12                                  12 
##                           Pea Aphid                     Pond Wolf Spider 
##                                  12                                  12 
##               Armoured Scale Family                     Diamondback Moth 
##                                  13                                  13 
##                       Eulophid Wasp                   Monarch Butterfly 
##                                  13                                  13 
##                       Predatory Bug                 Yellow Fever Mosquito 
##                                  13                                  13 
##                        Corn Earworm                    Green Peach Aphid 
##                                  14                                  14 
##                           House Fly                           Ox Beetle 
##                                  14                                  14 
##                  Red Scale Parasite                   Spined Soldier Bug 
##                                  14                                  14 
##               Western Flower Thrips Hemlock Woolly Adelgid Lady Beetle 
##                                  15                                  16 
##               Hemlock Wooly Adelgid                                Mite 
##                                  16                                  16 
##                         Onion Thrip                Araneoid Spider Order 
```

```
##                           16                             17
##                    Bee Order                 Egg Parasitoid
##                           17                             17
##                 Insect Class     Moth And Butterfly Order
##                           17                             17
##  Oystershell Scale Parasitoid     Black-spotted Lady Beetle
##                           17                             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
##                 Codling Moth    Flatheaded Appletree Borer
##                           19                             20
##          Horned Oak Gall Wasp            Leaf Beetle Family
##                           20                             20
##            Potato Leafhopper    Tooth-necked Fungus Beetle
##                           20                             20
##               Argentine Ant                          Beetle
##                           21                             21
##                    Mason Bee                       Mosquito
##                           22                             22
##             Citrus Leafminer                Ladybird Beetle
##                           23                             23
##            Spider/Mite Class            Tobacco Flea Beetle
##                           24                             24
##                 Chalcid Wasp       Convergent Lady Beetle
##                           25                             25
##               Stingless Bee           Ground Beetle Family
##                           25                             27
##           Rove Beetle Family                 Tobacco Aphid
##                           27                             27
##                Scarab Beetle                 Spring Tiphia
##                           29                             29
##                  Thrip Order        Ladybird Beetle Family
##                           29                             30
##                   Parasitoid                 Braconid Wasp
##                           30                             33
##                 Cotton Aphid                Predatory Mite
##                           33                             33
##          Sweetpotato Whitefly                  Aphid Family
##                           37                             38
##               Cabbage Looper          Buff-tailed Bumblebee
##                           38                             39
##               True Bug Order       Sevenspotted Lady Beetle
##                           45                             46
##                 Beetle Order  Snout Beetle Family, Weevil
##                           47                             47
##           Erythrina Gall Wasp               Parasitoid Wasp
##                           49                             51
##        Colorado Potato Beetle                 Parastic Wasp
```

```
##                           57                          58
##          Asian Citrus Psyllid           Minute Pirate Bug
##                           60                          62
##             European Dark Bee                    Wireworm
##                           66                          69
##                Euonymus Scale           Asian Lady Beetle
##                           75                          76
##               Japanese Beetle            Italian Honeybee
##                           94                         113
##                    Bumble Bee         Carniolan Honey Bee
##                          140                         152
##          Buff Tailed Bumblebee              Parasitic Wasp
##                          183                         285
##                     Honey Bee                     (Other)
##                          667                         670
```

Answer: The six most commonly studies species in this dataset (other than the "Other" category) are honey bee, parasitic wasp, buff tailed bumblebee, carniolan honey bee, bumble bee, and italian honeybee. They might be of interest over other insects due to the fact that they are pollinators and their population dynamics might drastically influence the health of our ecosystems.

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?

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

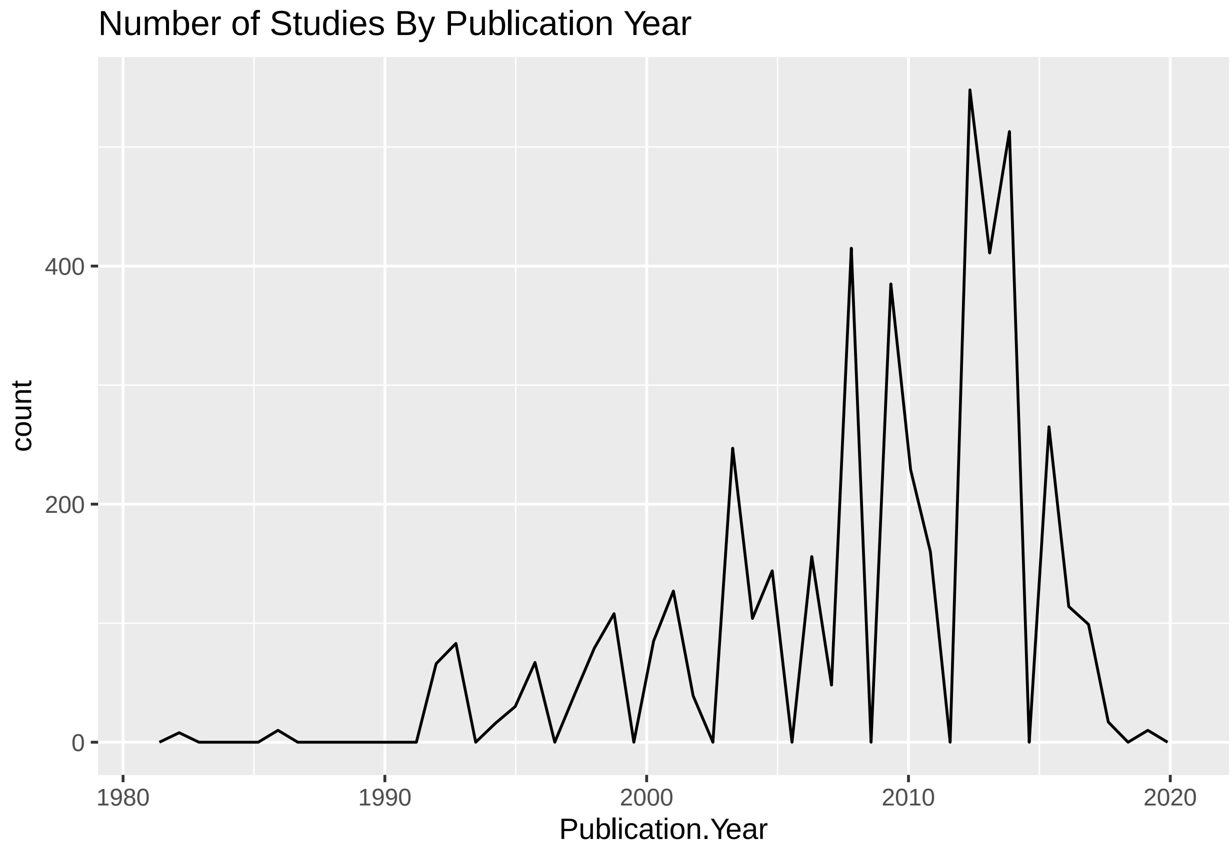
```
## [1] "factor"
```

Answer: The class is "factor". It is not numeric due to the fact that it is read in as strings(characters) since some cells have characters in them, and we had set stringsAsFactors to TRUE; and thus it was read in as factors.

Explore your data graphically (Neonics)

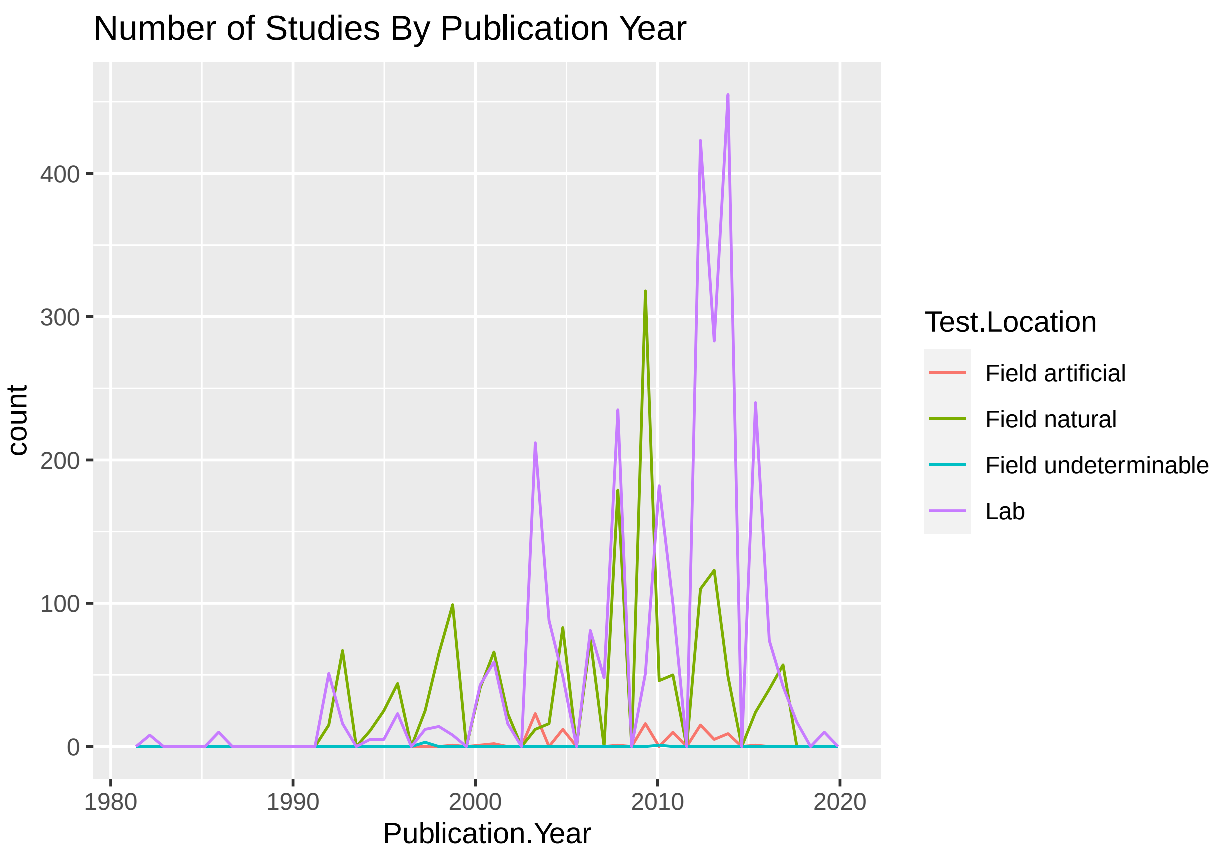
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50) + ggtitle('Number of Studies By Publication Year')
```



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

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 50) +
  ggtitle('Number of Studies By Publication Year') +
  theme(legend.position = "right")
```



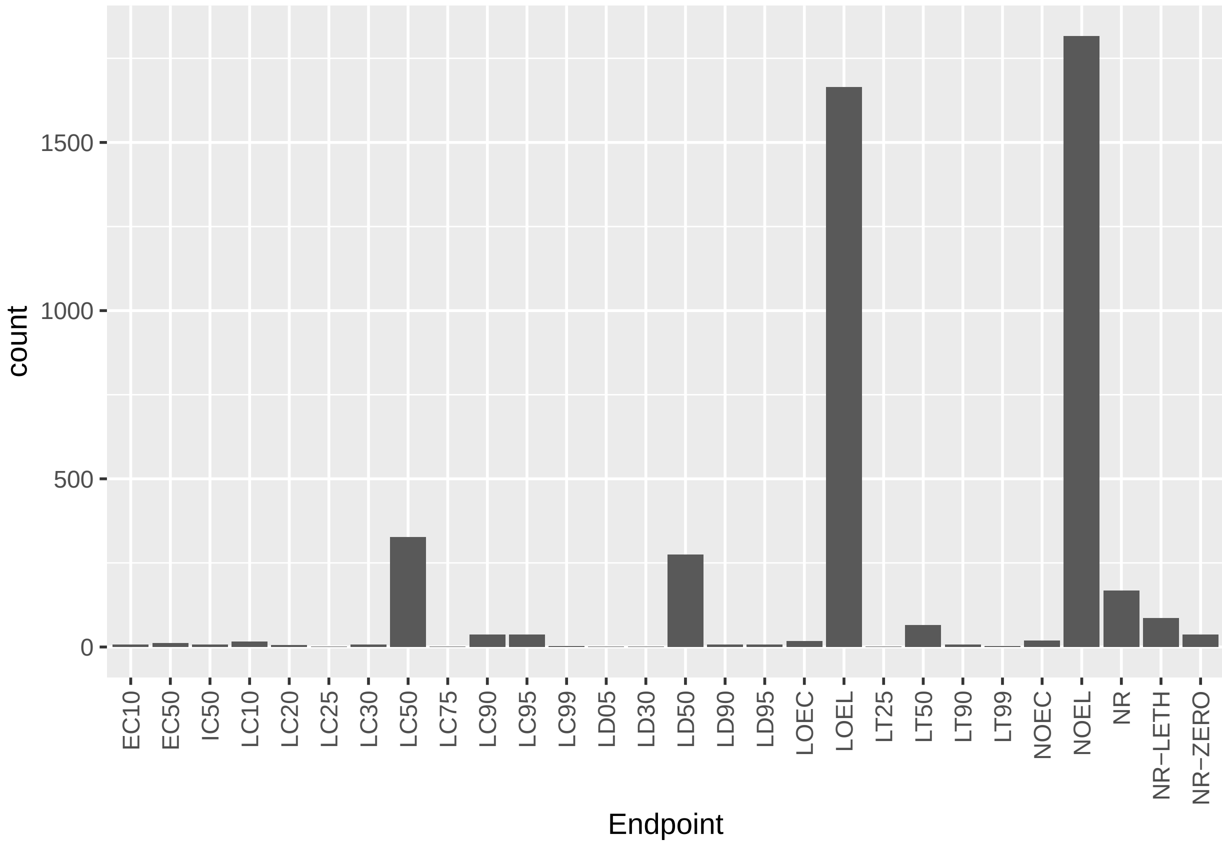
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations seem to be labs. The second most common test locations seem to be "field natural". While lab is the most common most of the time, field_natural surpassed lab in a few years.

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…]

```
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 is defined as "No-observable-effect-level: highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test (NOEAL/NOEC)". LOEL is defined as "Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC)".

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.

```
# check the class
class(Litter$collectDate)
```

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

```
# change the class of collectDate column
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")

# check the class again
class(Litter$collectDate)
```

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

```r
# which dates litter was sampled in August 2018
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`?

```r
length(unique(Litter$plotID))
```

```
## [1] 12
```

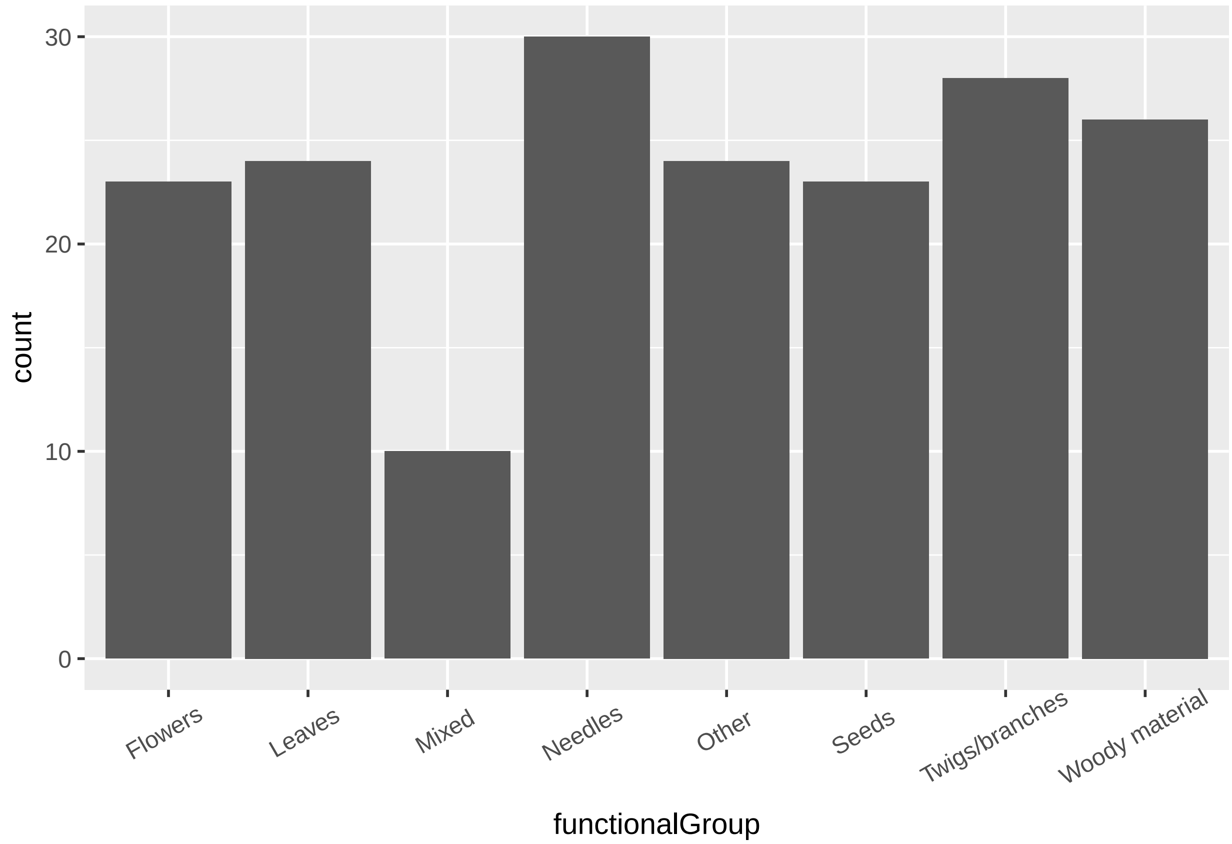
```r
summary(Litter$plotID)
```

```
## 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: 12 plots were sampled. unique() gives you all the unique instances that showed up in plotID. summary() gives you the counts of each instances.

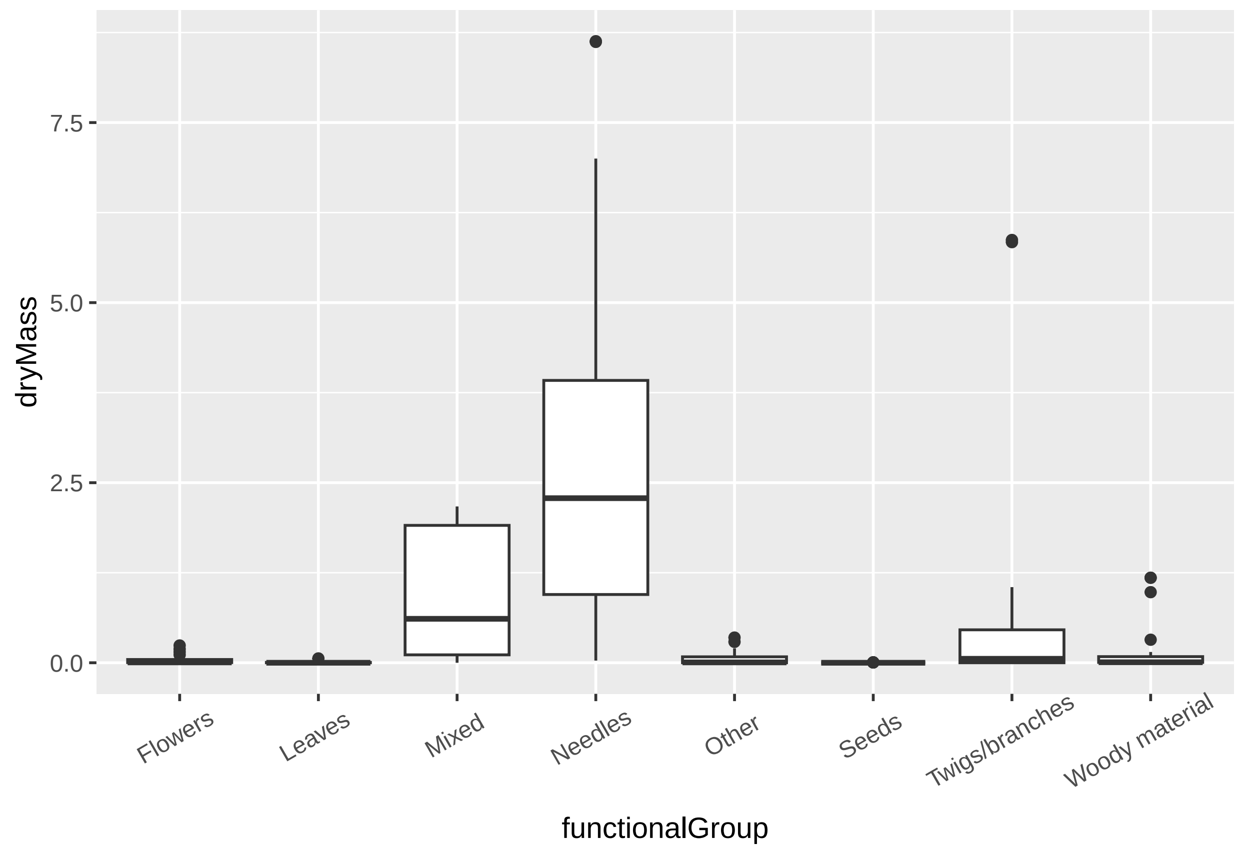
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 30, vjust = 0.8, hjust=0.6))
```

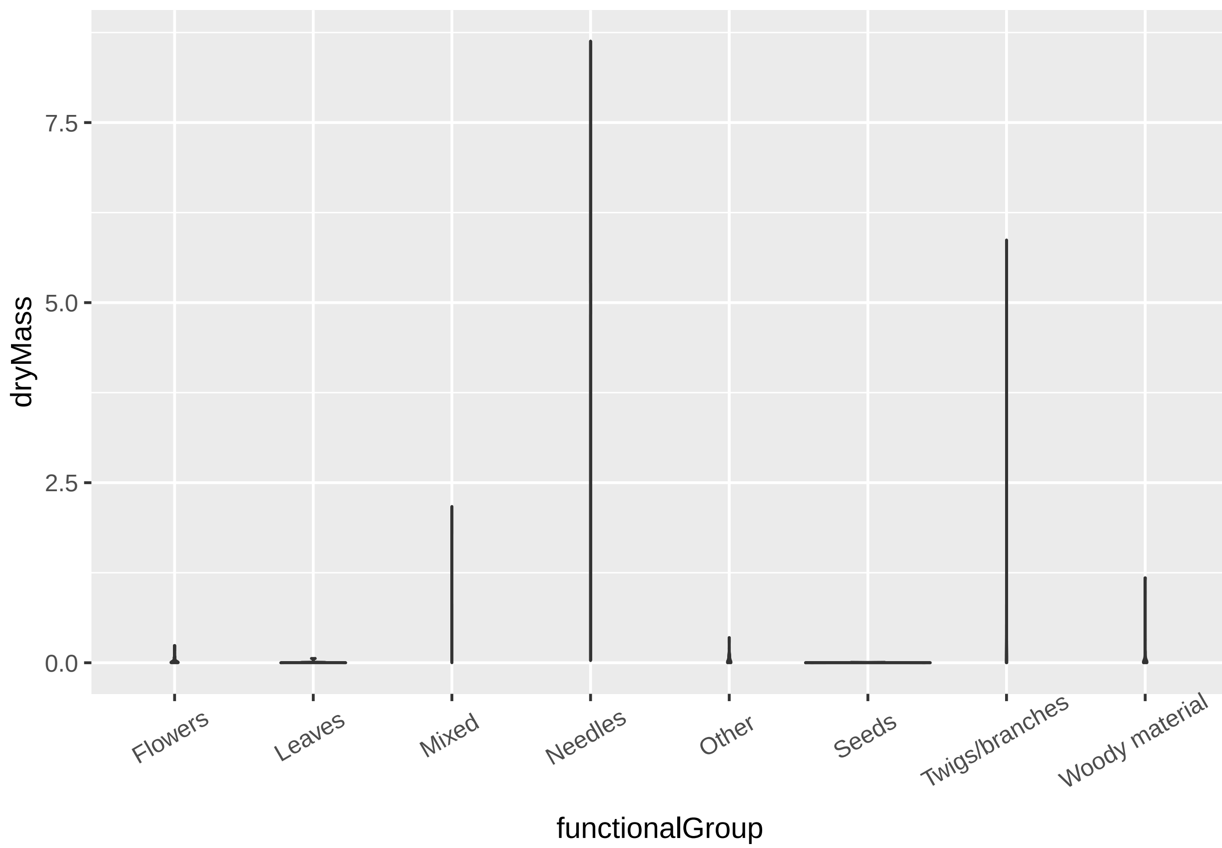


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass)) +
  theme(axis.text.x = element_text(angle = 30, vjust = 0.8, hjust=0.6))
```




```
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass), draw_quantiles = c(0.25, 0.5, 0.75)) +
    theme(axis.text.x = element_text(angle = 30, vjust = 0.8, hjust=0.6))
```



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

> Answer: In this case, boxplot is a more effective visualization option than violin plot because you can see the distribution more. For the violin plot, since the dry mass of needles generally was much higher than the rest of the categories, it is not obvious at all how they are distributed.

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

> Answer: Needles tend to have the highest biomass.