Assignment 3: Data Exploration

Yilin Zhong

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)
getwd()
```

```
## [1] "C:/Users/victo/Desktop/Spring 2023/EDA/EDA-Spring2023"
```

```
Neonics<-read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",stringsAsFactors = T)
Litter<-read.csv("./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 are interested in the ecotoxicology of neonicotionoids on insects because neonicotionoids is a class of insecticides, which contains toxic chemicals to kill insects. We want to know the toxicity that this class of insecticide contains to determine their effects on the environment when people applied it in the fields. The ecotoxicology of neonicotionoids allow us to determine the amount people should use while have the least toxic effects on the environment and people.

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 are interested in studying litter and woody debris that falls to the ground in forests because they play an important role on carbon budget and nutrient cycling. They are a source of energy in the aquatic ecosystems. In addition, they provide habitat for terretrial and aquatic organisms. Moroever, they determines the structure and roughness of their area, which influence water flow and sediment transports.

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 sampled as part of the NEON network by elevated and ground traps, respectively. 1.Litter and fine woody debris sampling took place at terrestrial NEON sites that contain woody vegetation >2m tall. Together with most of NEON's plant productivity measurements, sampling for this product occurs only in tower plots. Locations of tower plots are choosen randomly within the 90% flux footprint of the primary and secondary airsheds. 2.One litter trap pair, which includes one elevated trap and one ground trap, is deployed for every 400 m^2 plot area, led to 1-4 trap pairs per plot. 3. Plot edges must be separated by a distance 150% of one edge of the plot; plot centers must be greater than 50 meter from large paved roads and plot edges must be 10m from two-track dirt roads; plot centers must be 50 meter from buildings and other non-NEON infrastructure; streams larger than 1 meter must not intersect plots.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) #dimension is 4623 observations and 30 variables
```

```
## [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) #counts the total numbers in each effect category
```

```
##     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 are mortality and population, which each exceeds 1000. These effects are speicifically be of interest because mortality reveals the measurements and endpoints where the cause of death is by direct action of the chemical. We want to know how many deaths are cause by the chemical and using these info to decide further action if needed. As for population, it reveals the measurements and endpoints related to a group of organisms or plants of the same species occupying the same area at a given time. We interested in population because we want to know what groups of organisms or plants are in the specific location at a sepcific time to see if there are exposure to chemicals.

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)) #counts the total numbers in each studied species
```

```
##                          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 studied species in the dataset are Honey bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. These species are all certain types of bees. They might be of interest over other insects because they have a role of pollinators that carry large grains of pollen between plants. Since they carry pollens, they also carry the insecticides that we are interested in our study, which they can expose the chemicals to other species as they fly around. This can work the other way as well, as they fly around, they can be easily exposed to the chemicals as they often have close contacts with plants.

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.) #factor class
```

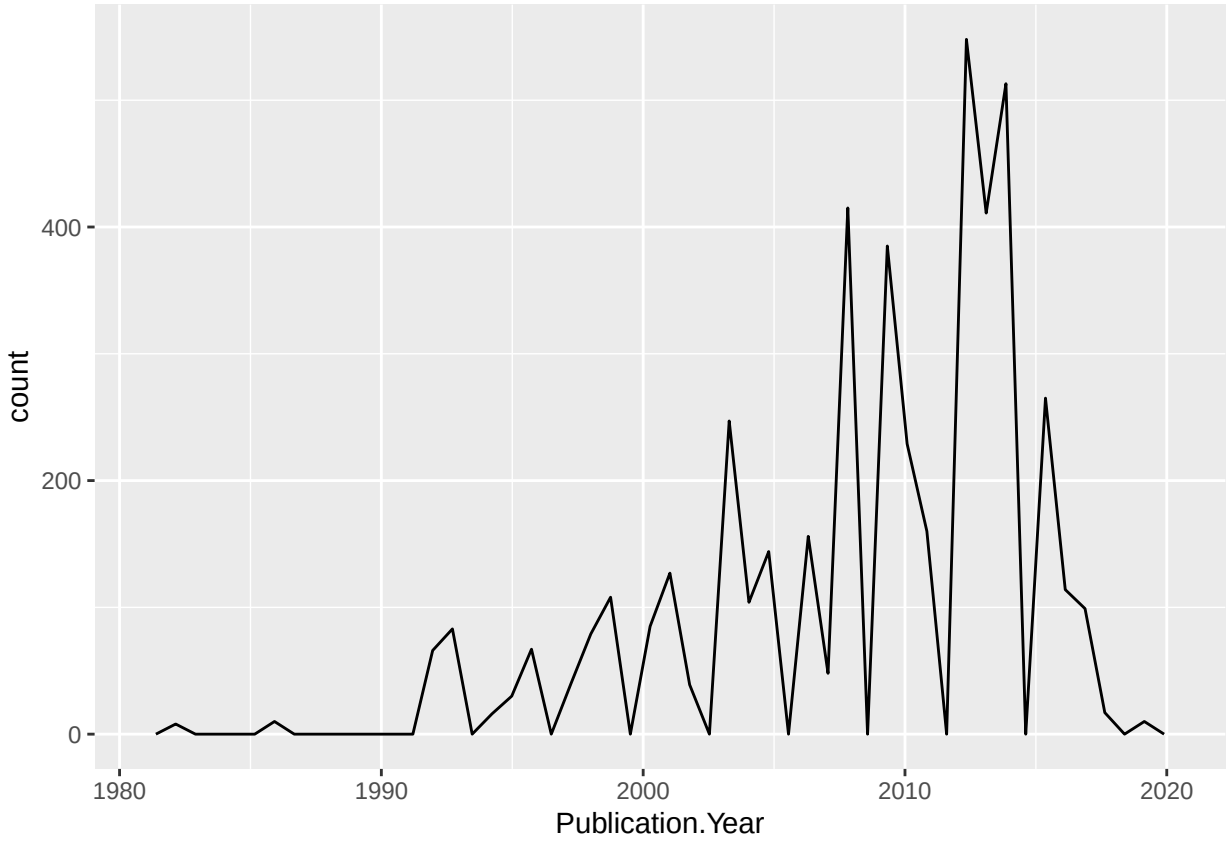
```
## [1] "factor"
```

Answer:The class for this column in the dataset is factor. It is not numeric because not all the values are in numbers. For instance, it contains NR, which means the number is not reported. Numeric class is not the best class to use in this case.

Explore your data graphically (Neonics)

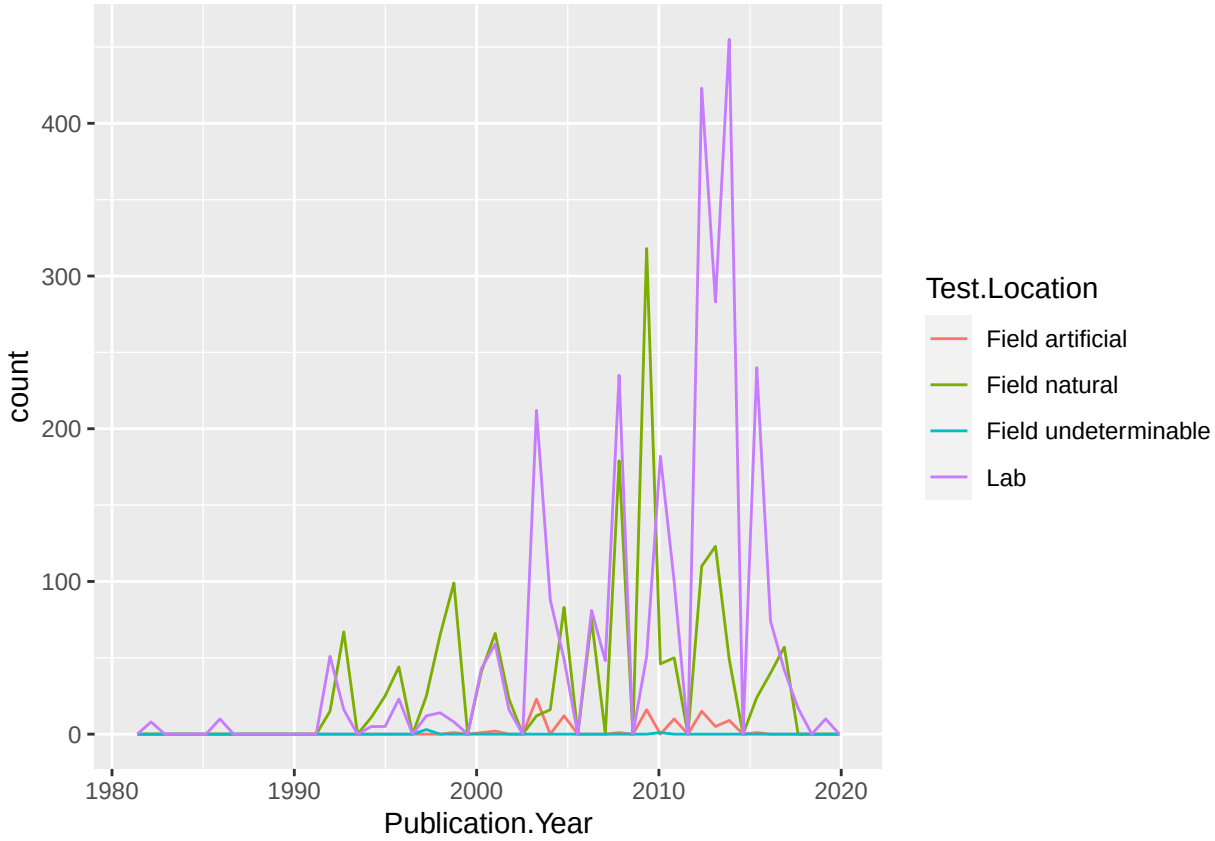
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
#plot of the number of studies conducted by publication year
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50)
```



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

```
#plot of the number of studies conducted by publication year with color aesthetic that shows
#different Test.Location and displayed as different colors.
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color= Test.Location), bins = 50)
```



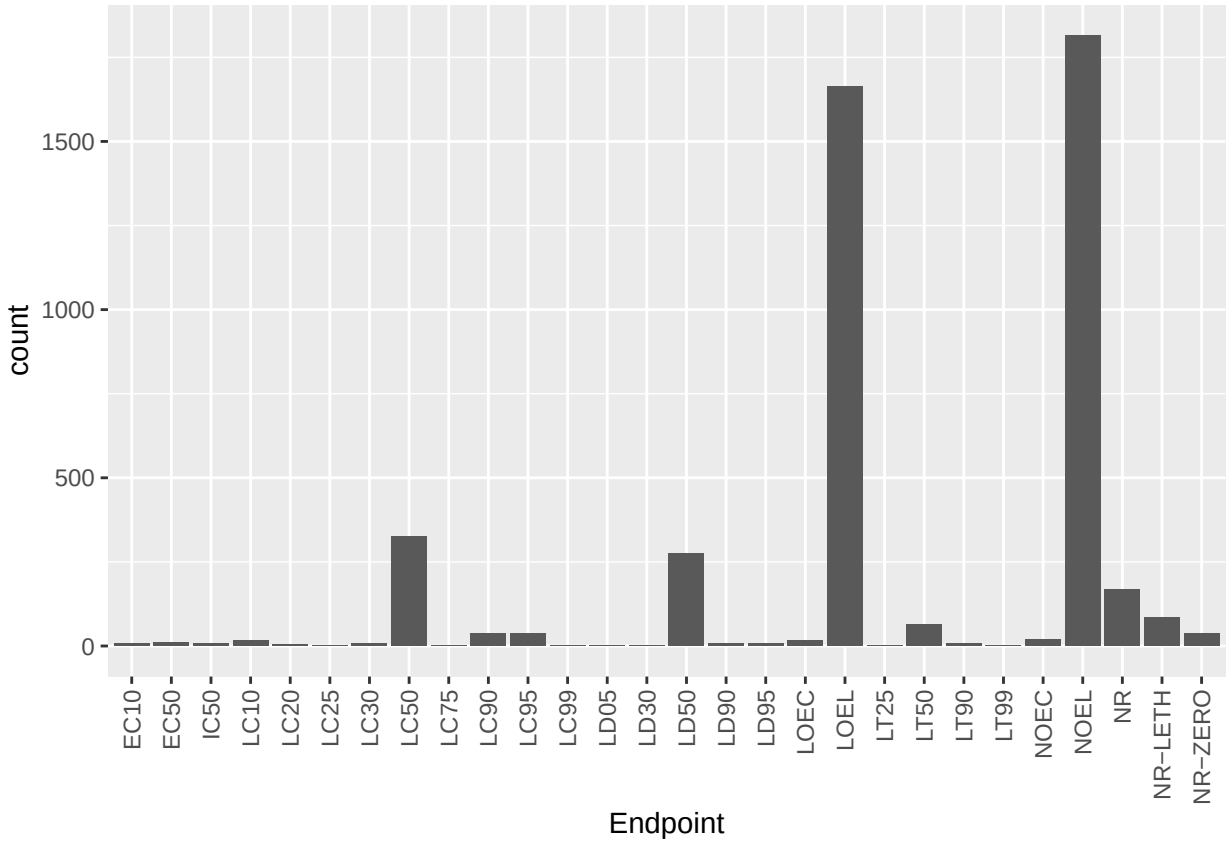
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common locations are lab and field nature. These two locations remain the most common locations at all time. However, in some years, field nature exceeds lab as the most common location, such as from year 1997 to 2000, and 2008 to 2010. For the rest of the timeframe, lab is the most common location.

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...]

```
#bar graph of 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 endpoints are LOEL and NOEL. LOEL is the lowest observable effect level, which means the lowest dose (concentration) producing effects that were significantly different from responses of controls. NOEL is the no observable effect level, which means the highest dose producing effects not significantly different from responses of controls according to the author's reported statisical test.

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) #the class is factor
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
class(Litter$collectDate) #now is date
```

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

```
unique(Litter$collectDate) # sampled on 2018-08-02 and 2018-08-30
```

```
## [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`?

```
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
```

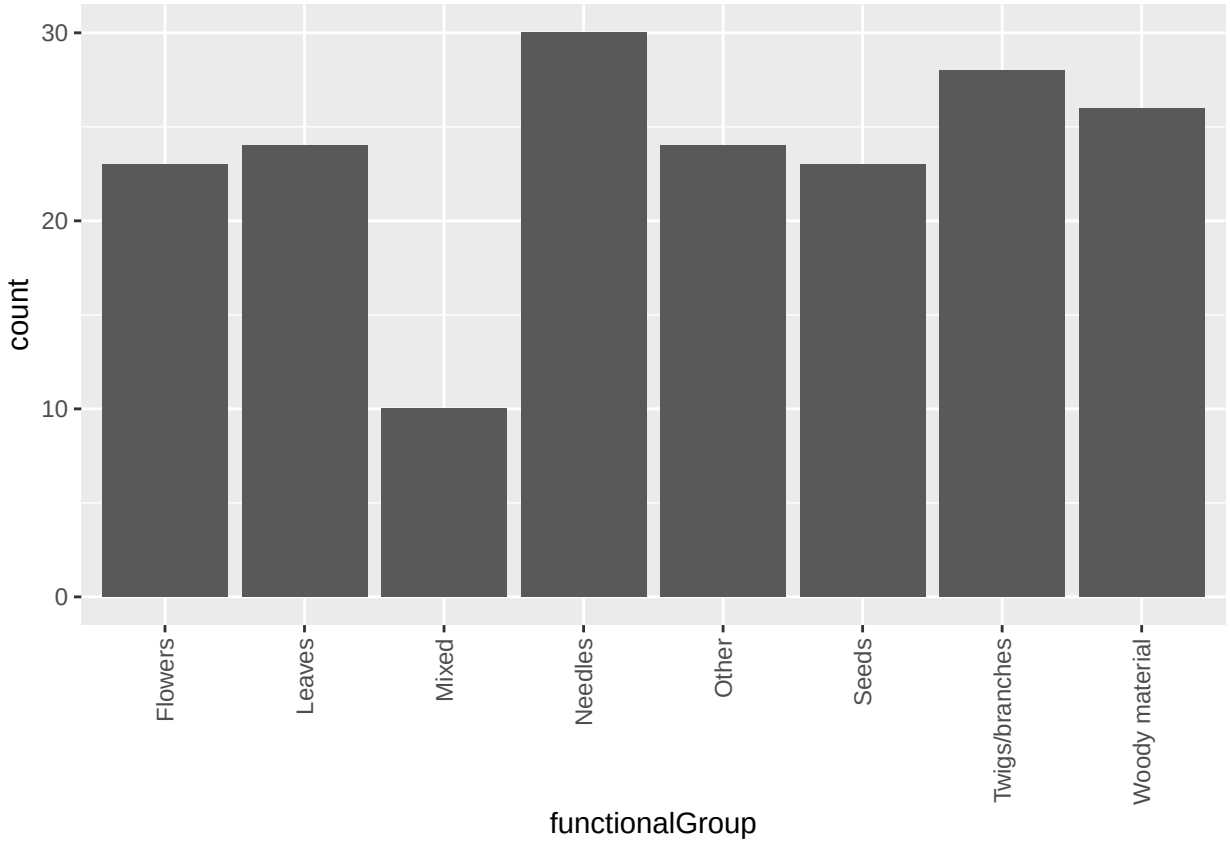
```
summary(Litter$plotID)
```

```
## 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: There are 12 plots sampled at Niwot Ridge. Unique function eliminate duplicate values, which shows how many unique plots are sampled at Niwot Ridge. As for summary, it shows how many collections took place at each plot.

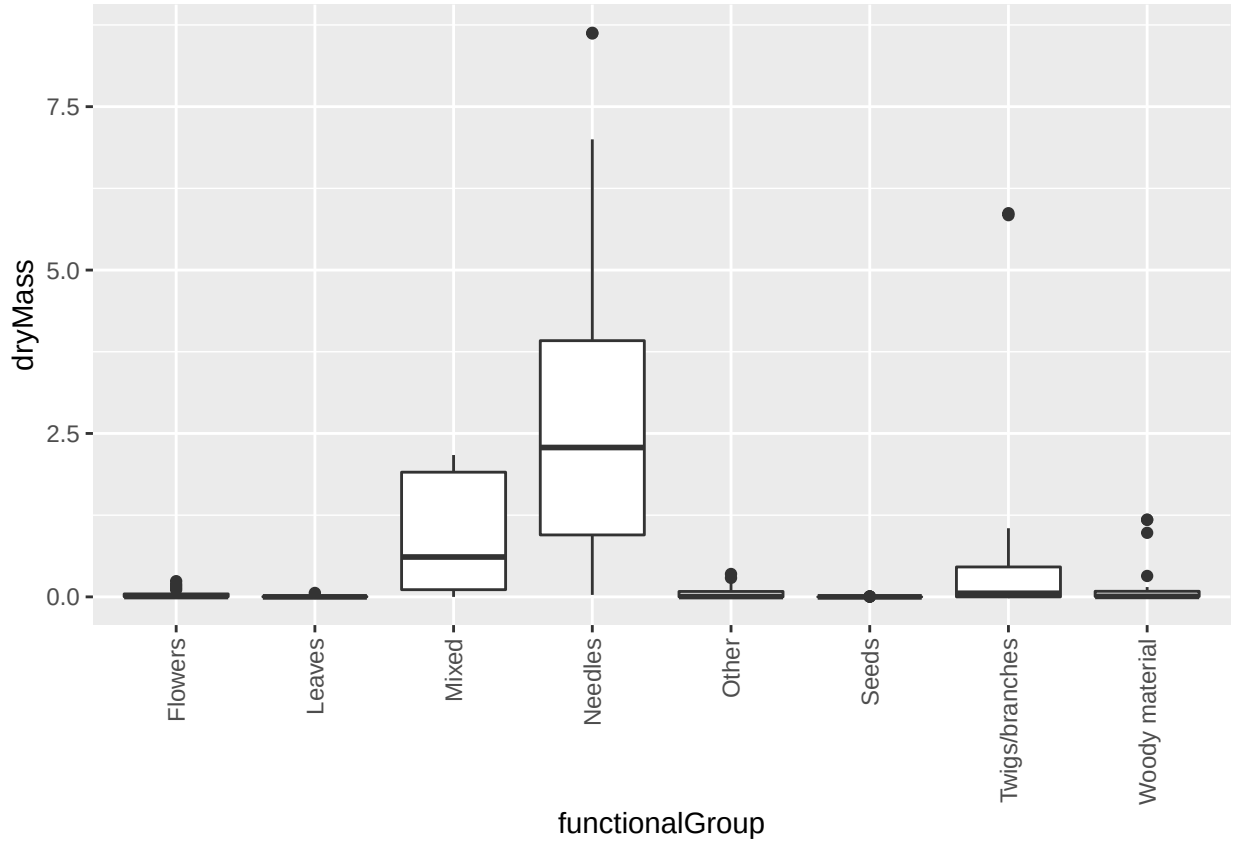
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
# bar graph of functionalGroup counts
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar()+
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



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

```
#boxplot of dryMass by functionalGroup and we rotate the x label for better orientation
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

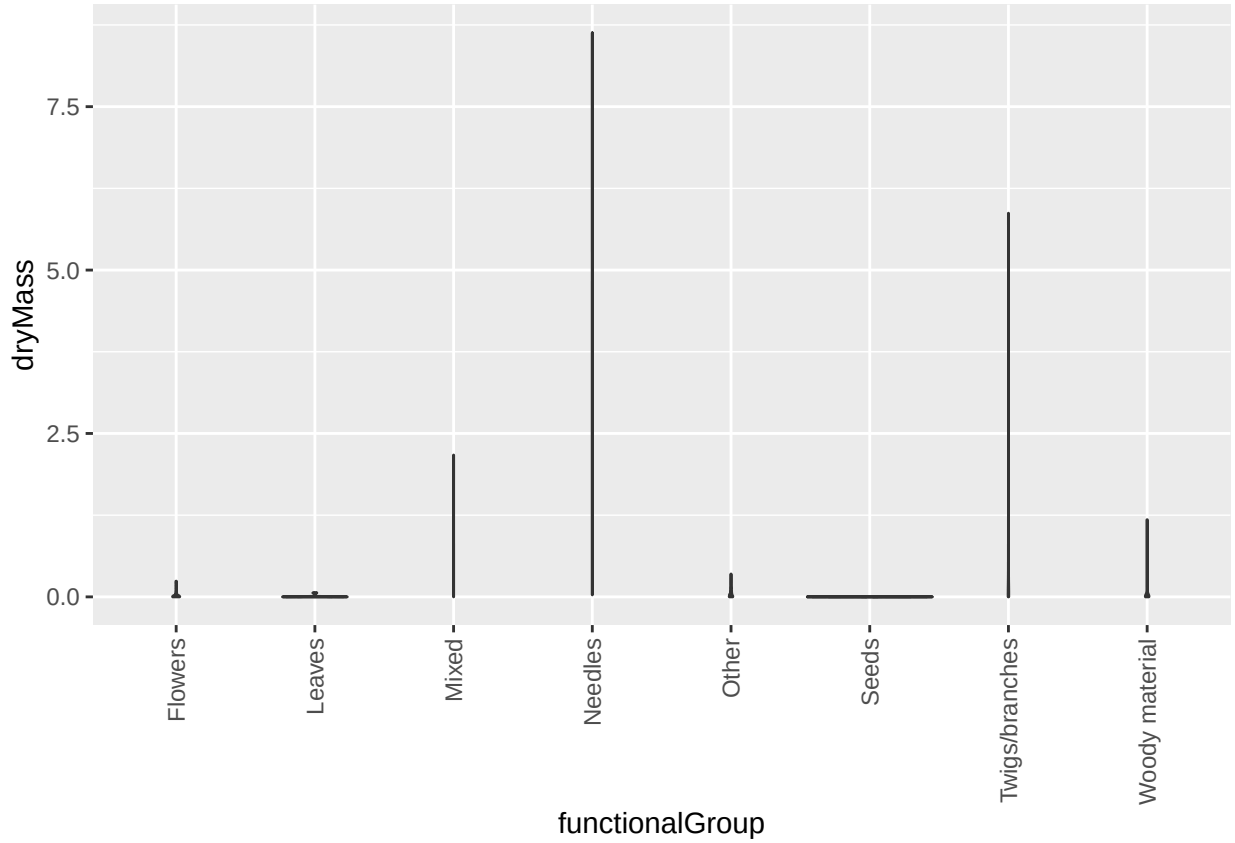



```
#violin plot of dryMass by functionalGroup and we rotate the x label for better orientation
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75))+
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```



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

> Answer:Boxplot is a more effective visualization option than the violin plot in this case because the sample size is relatively small, the density estimations that violin plot usually shows does not work in here. It does not have enough data to show the density estimations.

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

> Answer:Needles have the highest biomass at these sites, and Mixed have the second highest biomass.