Assignment 3: Data Exploration

Zhaoxin Zhang

Fall 2024

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); library(ggplot2)

#Check work directory
here()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```r
setwd(here())
#Upload the two datasets
Neonics <- read.csv(
  file = here('Data','Raw','ECOTOX_Neonicotinoids_Insects_raw.csv'),
  stringsAsFactors = T
)

Litter <- read.csv(
  file = 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: Because neonicotinoids can be risky. They can be not only toxic to insects but also human beings and food. Since neonicotinoids are widely used in agriculture, it is important to know if these neonicotinoids are effective. According to the report from NRDC: https://www.nrdc.org/stories/neonicotinoids-101-effects-humans-and-bees, neonicotinoids can hurt the ecosystem, so it is important to study on it.

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: According to the report from USDA: https://www.srs.fs.usda.gov/pubs/gtr/gtr_srs038/gtr_srs038-scheungrab001.pdf litter and woody debris is important because it can store carbon, recycles nutrients and plays an essential role for aquatic ecosystems. Litter and woody debris is biomass, which can be a type of renewable energy.

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.Randomly selected locations of tower plots 2.The presentage of vegetation covered is also important becuase it determine whether the plots should be targeted or randomized 3.Frequency for elevated traps is salient for Temporal Sampling Design.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
dim(Neonics)
```

```
## [1] 4623   30
```

```r
#There are 4623 rows and 30 columns
```

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...]

```r
sort(summary(Neonics$Effect))
```

```
##       Hormone(s)        Histology       Physiology          Cell(s) 
##                1                5                7                9 
##     Biochemistry     Accumulation     Intoxication    Immunological 
##               11               12               12               16 
##       Morphology           Growth        Enzyme(s)         Genetics 
##               22               38               62               82 
##        Avoidance      Development     Reproduction Feeding behavior 
##              102              136              197              255 
##         Behavior        Mortality       Population 
##              360             1493             1803 
```

Answer: The most common effects that are studied are "Population" and "Mortality", which are bigger than 1,000. These two effects show that researchers are more interested in how the neonicotinoids affect insects' population and mortality. They are interested in that becasue these two can directly show if the neonicotinoids decrease insects' population and increase their mortality. If neonicotinoids do, it means that they impact the growth of insects.

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...]

```r
summary(Neonics$Species.Common.Name) #Summarize the data
```

```
##                          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
```

```
help(summary) #Check what maxsum means
summary(Neonics$Species.Common.Name, maxsum = 7 )
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee      Italian Honeybee
##                   152                   140                   113
##               (Other)
##                  3083
```

```
#The question is asking for the six most commonly studied species
#We make the levels of 7
#because there would be one level shows "Other" besides the six most commonly studied species.
```

Answer: All of the six most commonly studied species are belong to the Apocrita suborder (wasps and bees). Studies show that neonicotinoids not only kill pests but also many other kinds of insects, especially bees. Therefore, researchers might want to know how the neonicotinoids can hurt bees and wasps, which are not targeted pest but are important to the ecosystem and food chain.

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...]

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

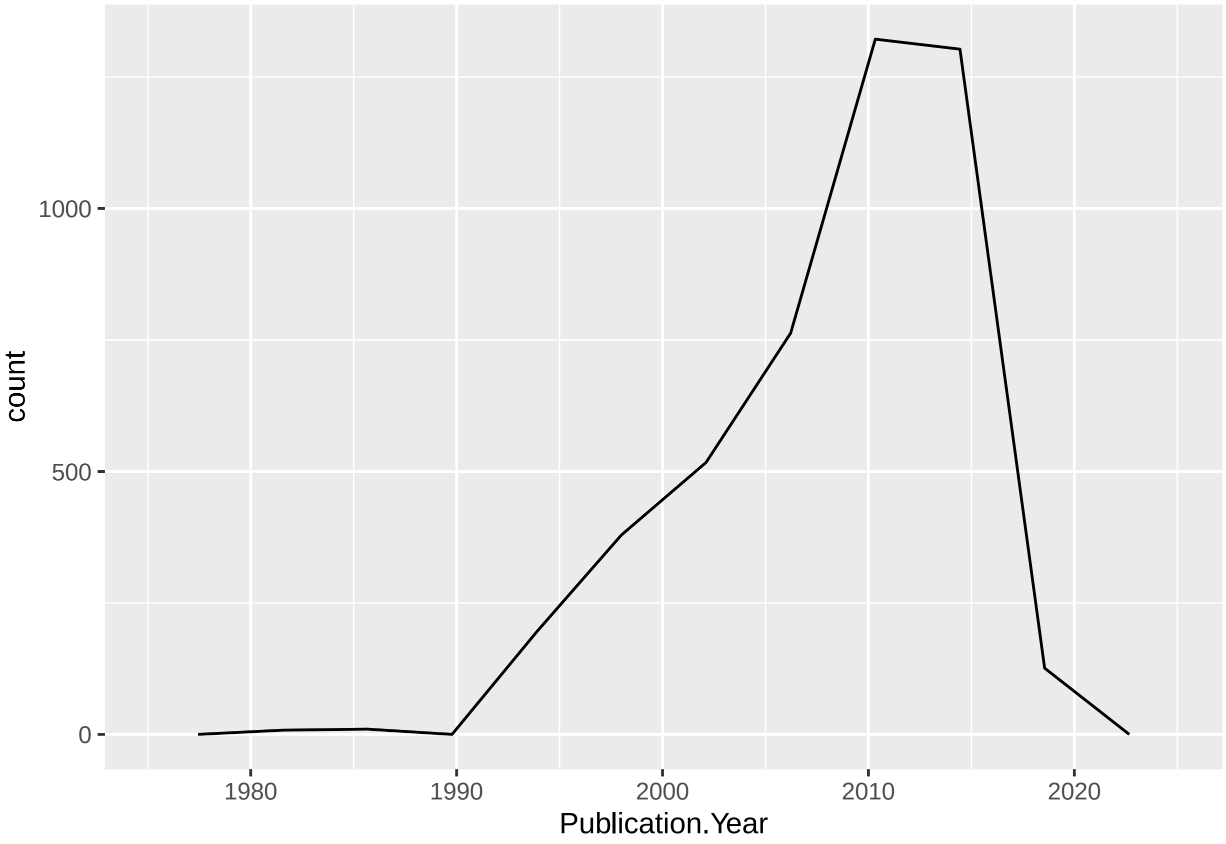
```
## [1] "factor"
```

> Answer: The "Conc.1..Author." is classified as factor. It is not numeric becasue it includes ranges such as smaller/bigger than specific numbers instead of just numbers, and it includes some special figures such as "<""\>" "/" which are not numeric.

Explore your data graphically (Neonics)

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

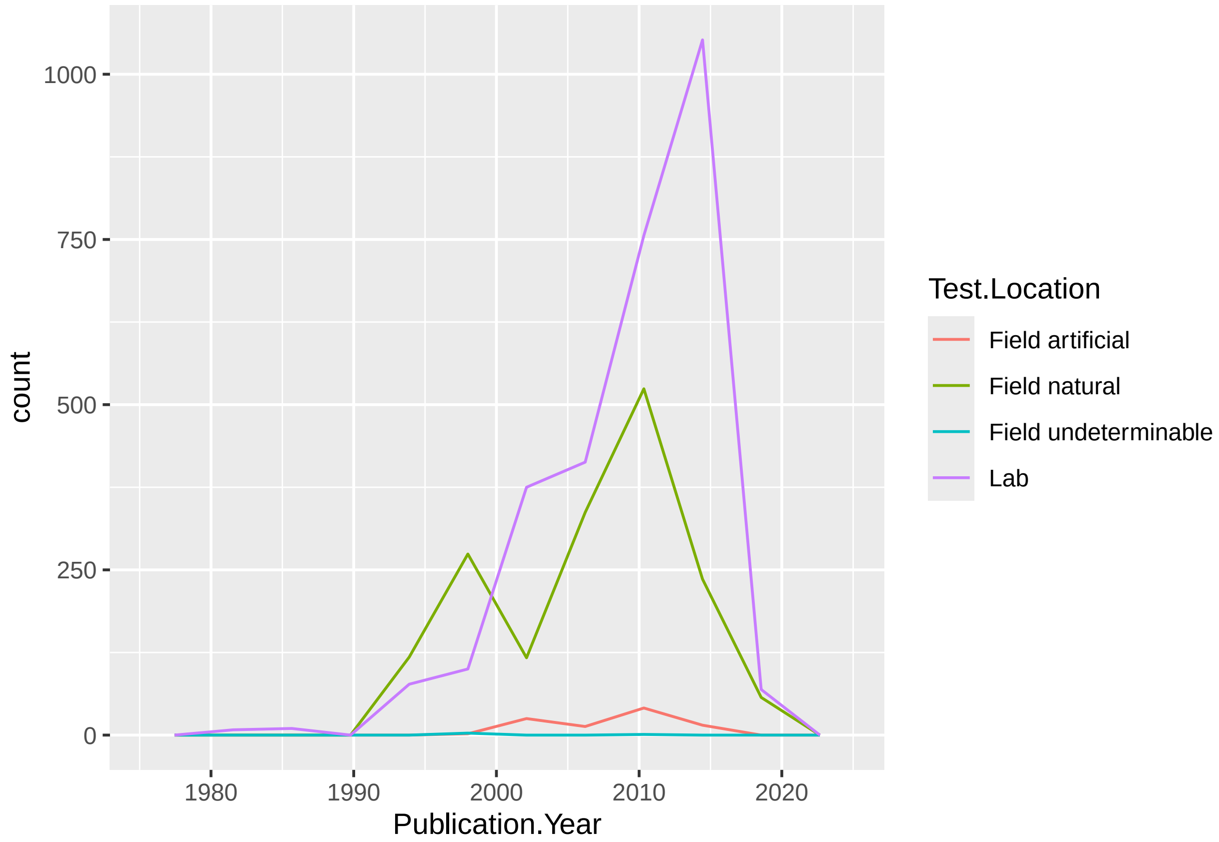
```
ggplot(Neonics)+
  geom_freqpoly(aes(x=Publication.Year), bins = 10)
```



```
#The bin is 10, showing the number of publications in around each three to five years.
```

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 = 10)
```



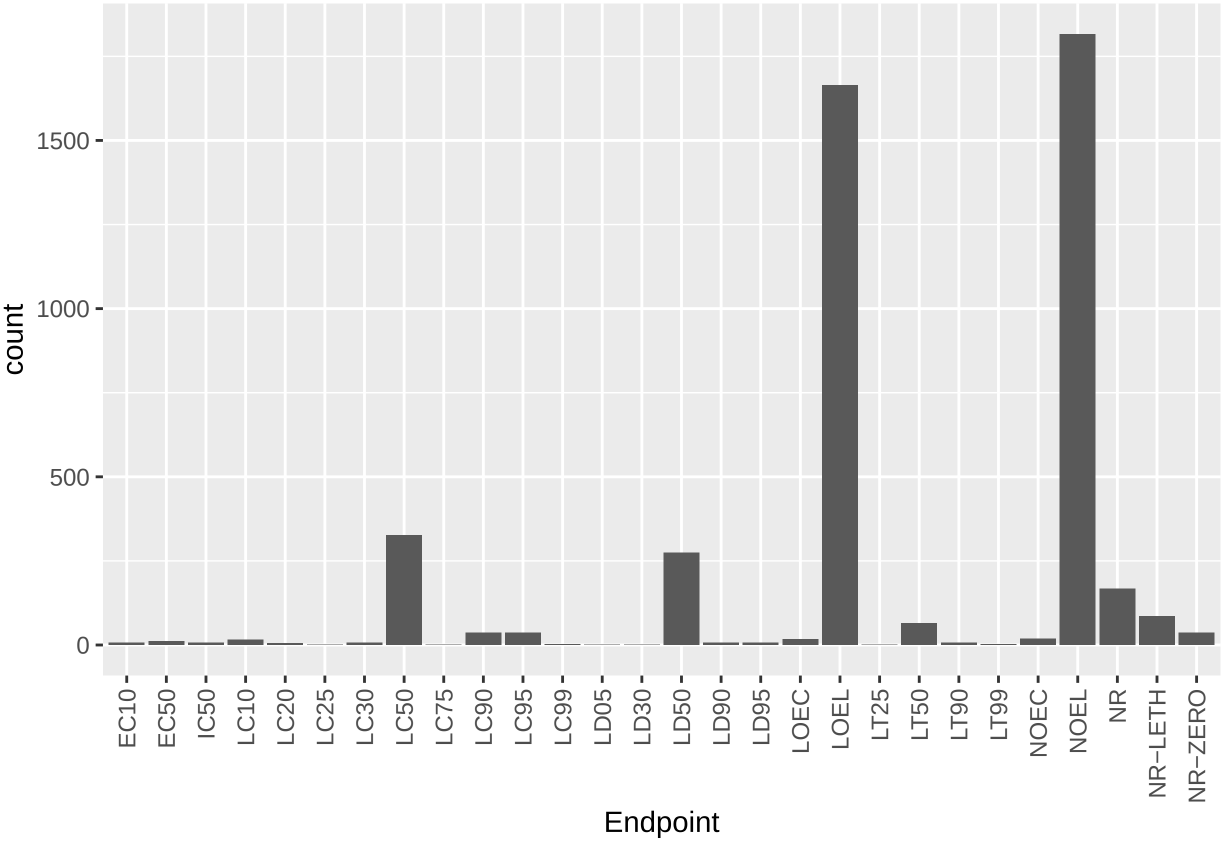
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are lab and field natural. The number of tests at field natural increased a lot between 2000s to 2010s, the number of tests at lab increased dramatically from 2005s to 2015s. From around 2010, the tests done at field natural decreased alot, and from around 2015, the tests done at lab dramatically decreased.

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 LOEL and NOEL. The LOEL represents the Lowest-observable-effect-level, which is the lowest dose (concentration) producing effects that were significantly different from responses of controls. The NOEL represents No-observable-effect-level, which is highest dose (concentration) producing effects not significantly different 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)
```

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

```
#It is a factor, not a data
Litter$collectDate <- as.Date(Litter$collectDate, format='%Y-%m-%d')
class(Litter$collectDate)
```

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

```
#Now it is a date
unique(Litter$collectDate)
```

```
## [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
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
```

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

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

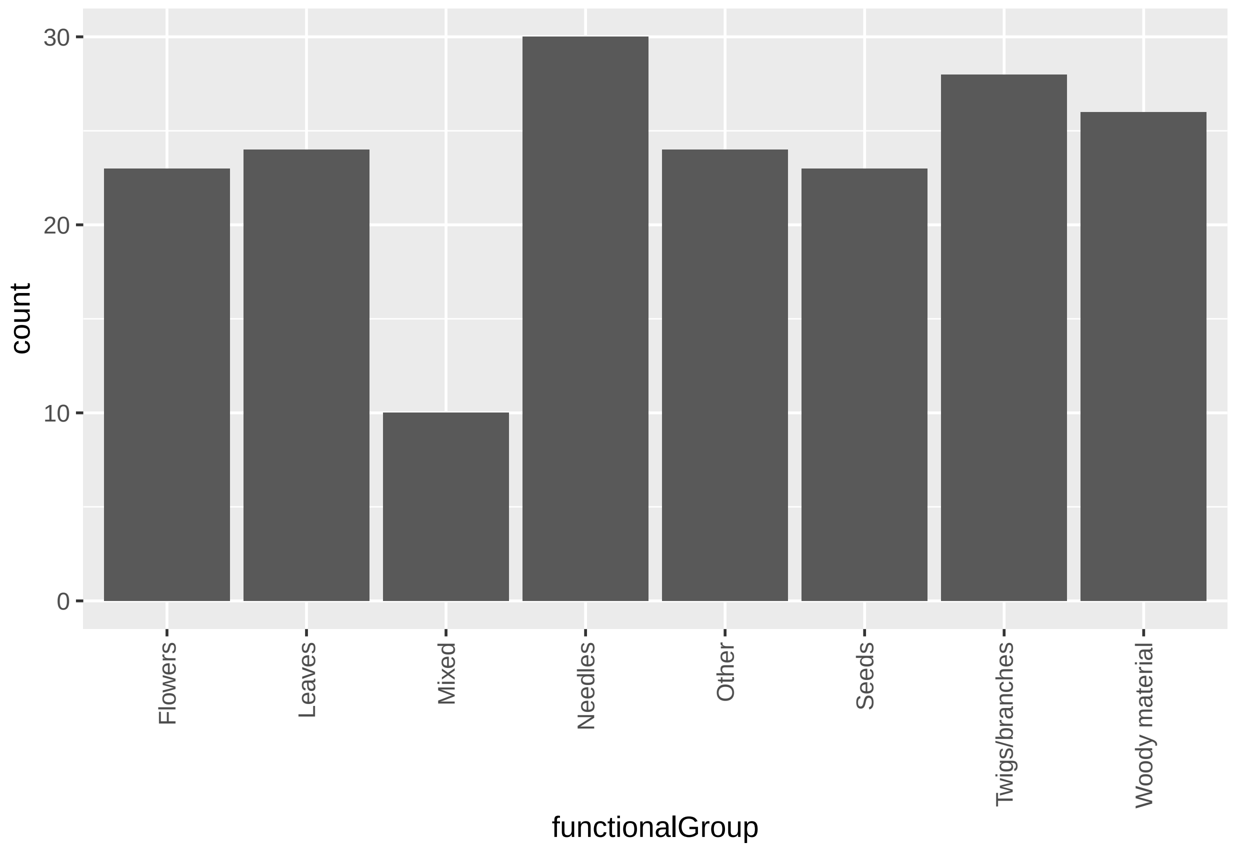
```r
length(summary(Litter$plotID))
```

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

Answer: The 'unique' only lists each plot's ID and how many different plots, without showinbg other information, while the 'summary'shows not only each plot's ID but also the number of each plot. However, the summary does not directly give the levels of the plot's ID, we need to count it one by one. If we use 'length', we can directly see how many different plots were sampled.

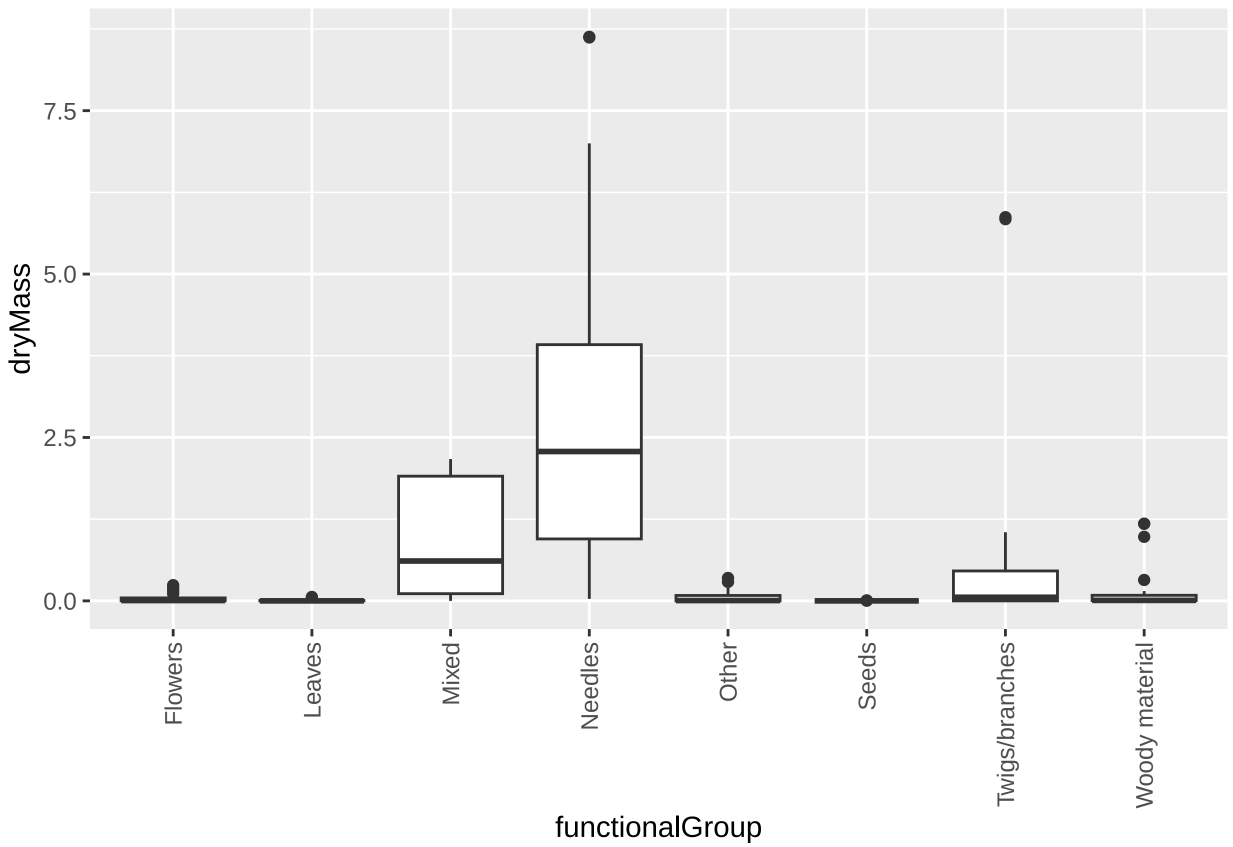
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
ggplot(Litter, aes(x=functionalGroup)) + geom_bar()+
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

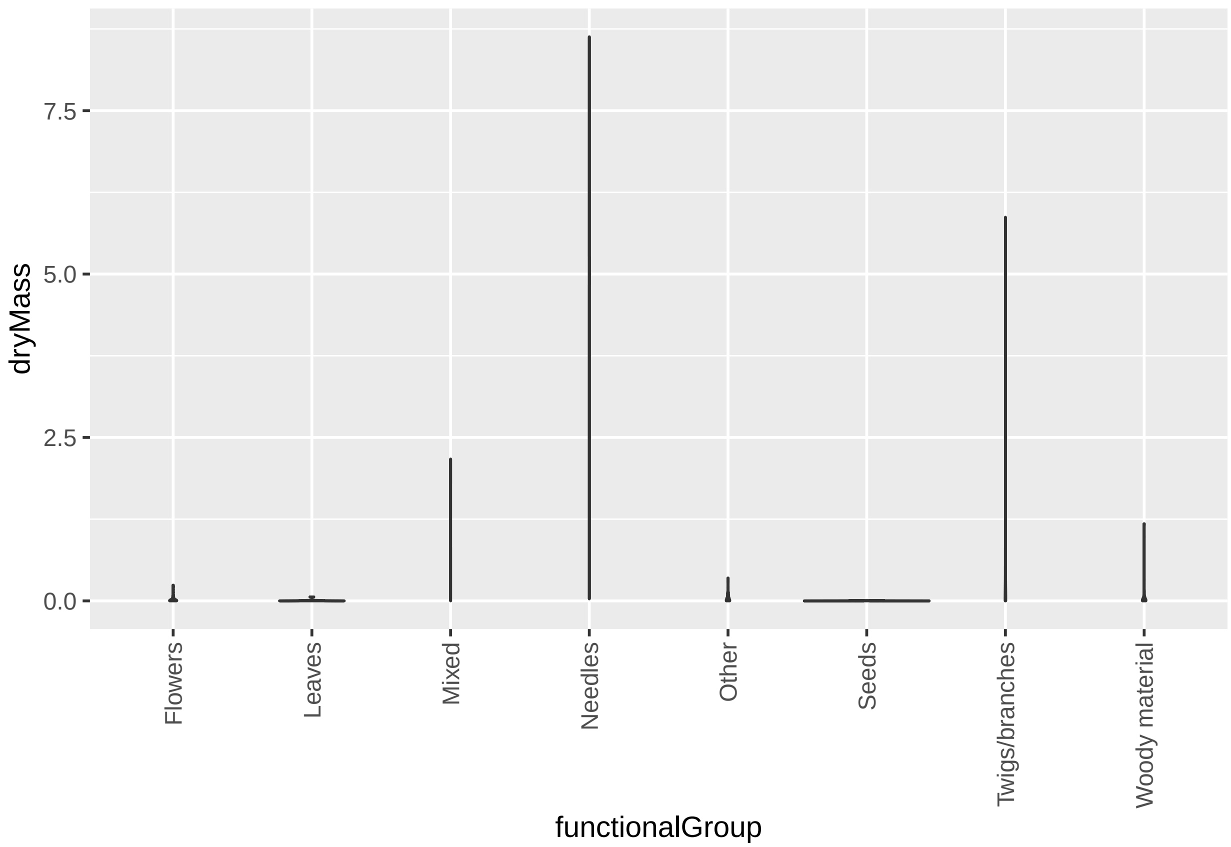


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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



```
ggplot(Litter, aes(y=dryMass, x=functionalGroup)) +geom_violin()+
    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: Because most of the functional groups' dry mass is smaller than 1. while some of the groups have dry mass larger that 5, the medians of functional groups are all smaller than 2.5. A boxplot can clearly show the median, range, and outliers. While the violin plot cannot show the median and only shows the range.

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

> Answer: Needless tends to have the highest biomass at these sites.