Assignment 3: Data Exploration

Aidan Power

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.

```
library(tidyverse) #Installing the relevant packages
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(here)
```

```
## here() starts at /home/guest/EDE_Fall2024
```

```
getwd() #Making sure my directory still connects to Git
```

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

```
here()
```

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

```
Neonics <- here('Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv')
print(Neonics)
```

```
## [1] "/home/guest/EDE_Fall2024/Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"
```

```
Neonics <- read.csv(
  file = here('Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv'),
  stringsAsFactors = T
)

Litter <- here('Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv')
print(Litter)
```

```
## [1] "/home/guest/EDE_Fall2024/Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv"
```

```
Litter <- read.csv(
  file = here('Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv'),
  stringsAsFactors = T
) #Loaded necessary packages and added the required datasets
```

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 interested in the ecotoxicology of neonicotoids based on how they effect pollinating insects. If insects that aid in crop reproduction die, there could be less yield of food.

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 studying litter and woody debris based on how it plays a very important role in carbon storage and nutrient cycling.

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. Trap placement is both targeted and randomized depending on vegetation cover and height. 2. 1-30 m^2 nested subplots utilized in 400 or 1600 m^2 plots. 3. All data taken at sites where there is evidence of woody vegetation being at least 2m tall.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

```
#Gave me the number of rows in columns listed in the console pane (4623 rows, 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...]

```
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
```

```
Common_Effects <- sort(summary(Neonics$Effect))
Common_Effects #Used sorted data to pick two most common effects
```

```
##       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: Population and mortality may be of the highest interest based on how they represent the efficacy of the insecticides being used, as lower population and higher mortality represent an effective insecticide.

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...]

```
summary(Neonics$Species.Common.Name) #Gets me a list of all species freq.
```

```
##                          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
```

```
Common_species <- sort(summary(Neonics$Species.Common.Name, maxsum = 7),
                       decreasing = TRUE)
#Most common first
Common_species
```

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

Answer: The six most commonly studied species are Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. These species all appear to bee in the bee/wasp family (striped with stingers). They of much importance due to how vital they are in ecosystem services such as pollination and even making honey for three of them (we do not want them to die).

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...]

```
Conc_Class <- class(Neonics$Conc.1..Author.) #Reads off the class
Conc_Class
```

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

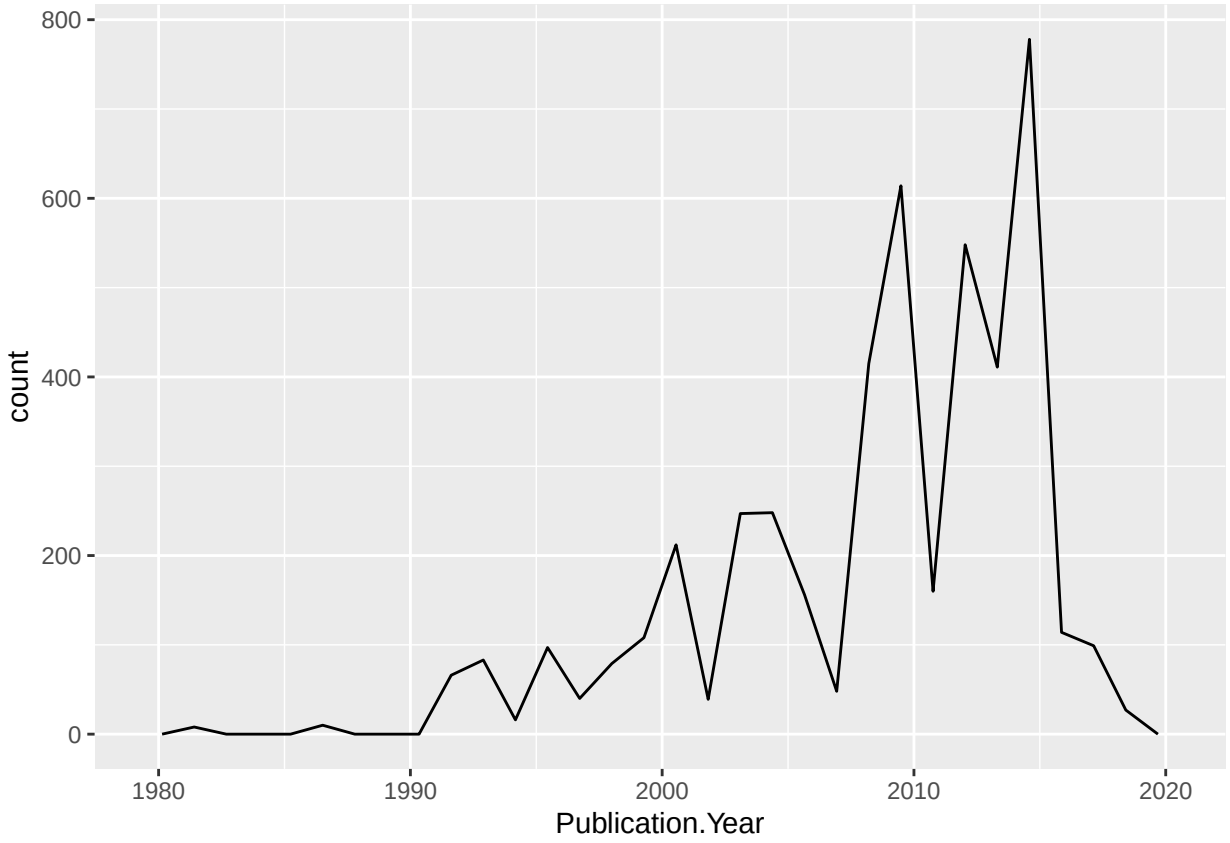
Answer: The class of "Conc.1..Author." is Factor. It likely is not numeric because of How it is a measurement (or categorical value) and not meant to be used in operations.

Explore your data graphically (Neonics)

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

```
help("geom_freqpoly")
ggplot(Neonics, aes(x = Publication.Year)) +
  geom_freqpoly() #Made graph based on dataset, set one column as x variable
```

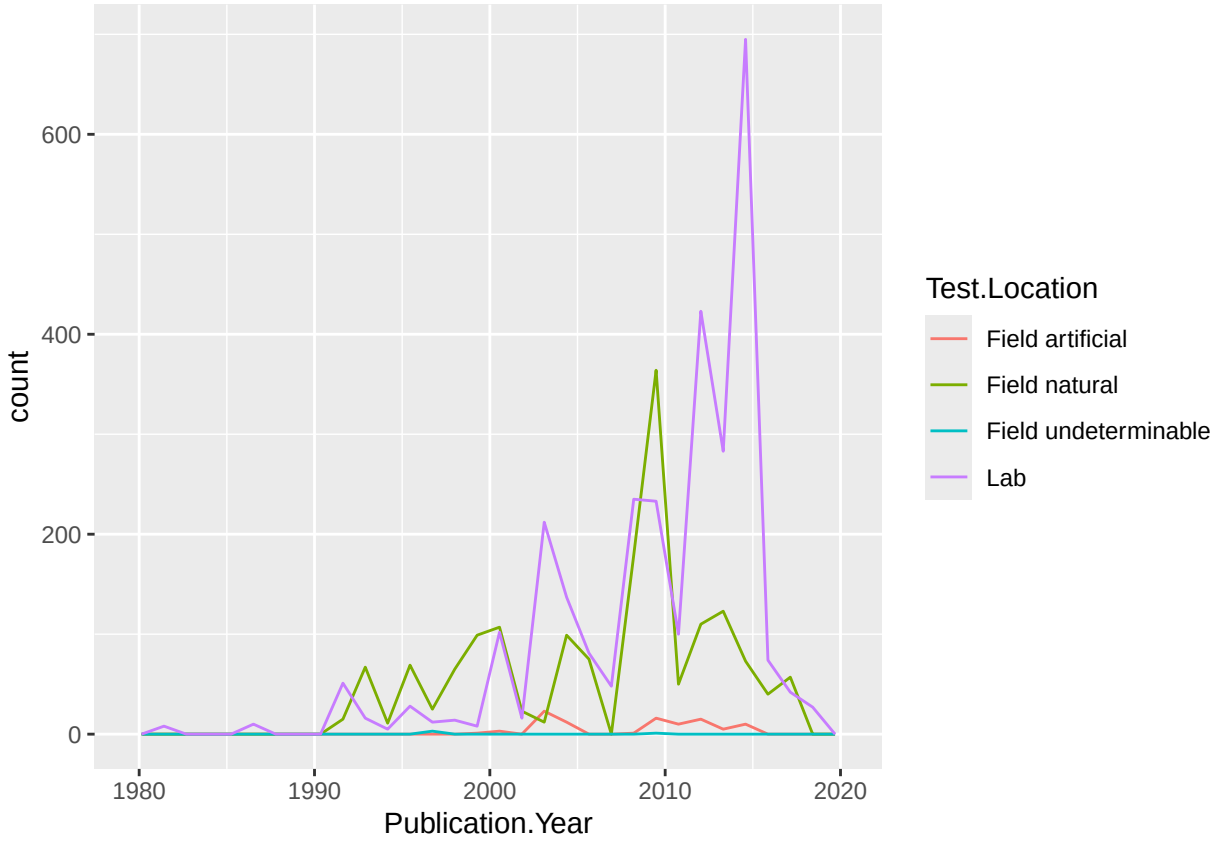
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

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

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



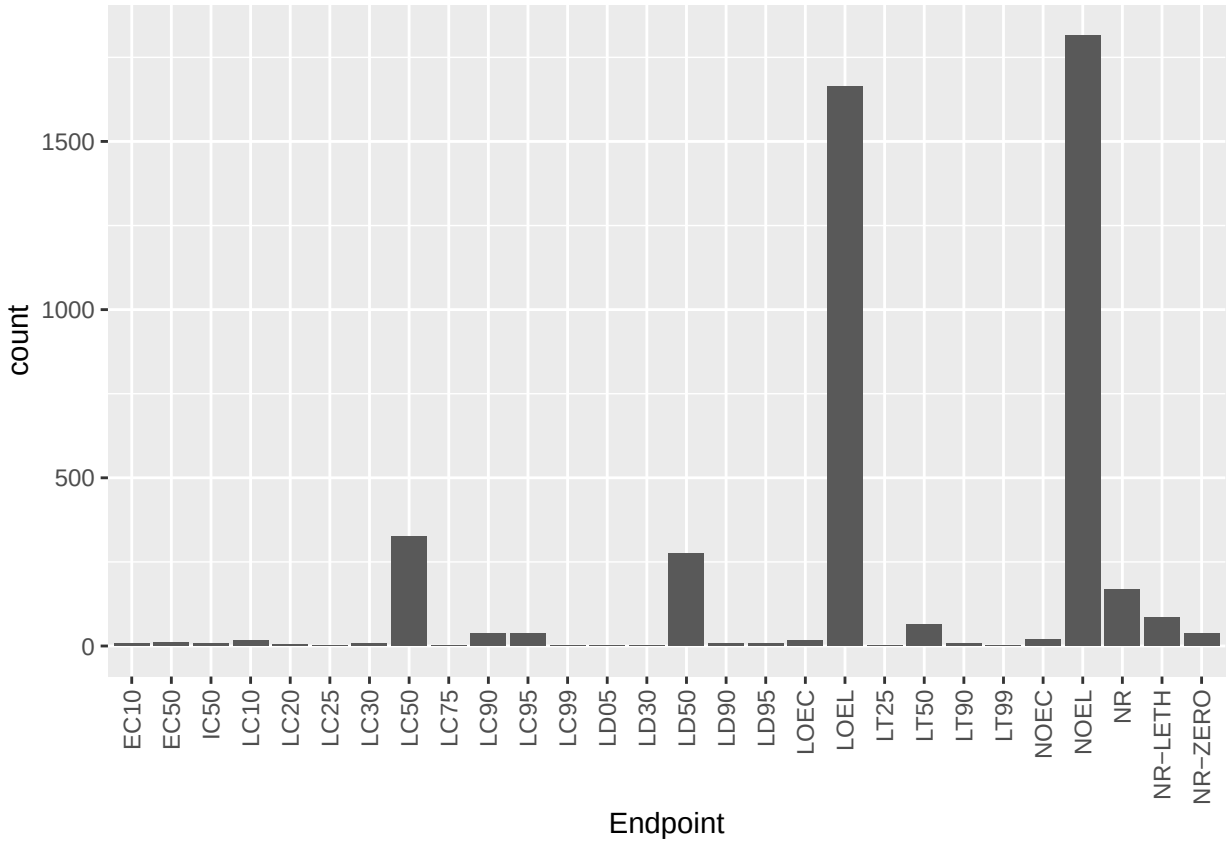
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: Lab and Field natural appear to be the most common test locations. Field natural is the most common in most of the 1990's and a few years around 2010, while Lab is the most common for most of the rest of the 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...]

```
help("geom_bar")
ggplot(Neonics, aes(x = Endpoint)) +
 geom_bar() +
 theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```




```
#Defined x as the endpoint column, used the provided code to rotate labels
```

Answer: NOEL and LOEL are the most common endpoints. LOEL is defined as "Lowest-observable-effect-level", which refers to a low dose producing significant effects. NOEL is defined as "No-observable-effect-level", which refers to the highest dose producing no significant effects.

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"
```

```
Litter$collectDate <- as_date(Litter$collectDate)
class(Litter$collectDate) #Changed the class to date and found two unique values
```

```
## [1] "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`?

```
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
```

```
length(unique(Litter$plotID)) #Counted number of unique values using length()
```

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

```
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
```

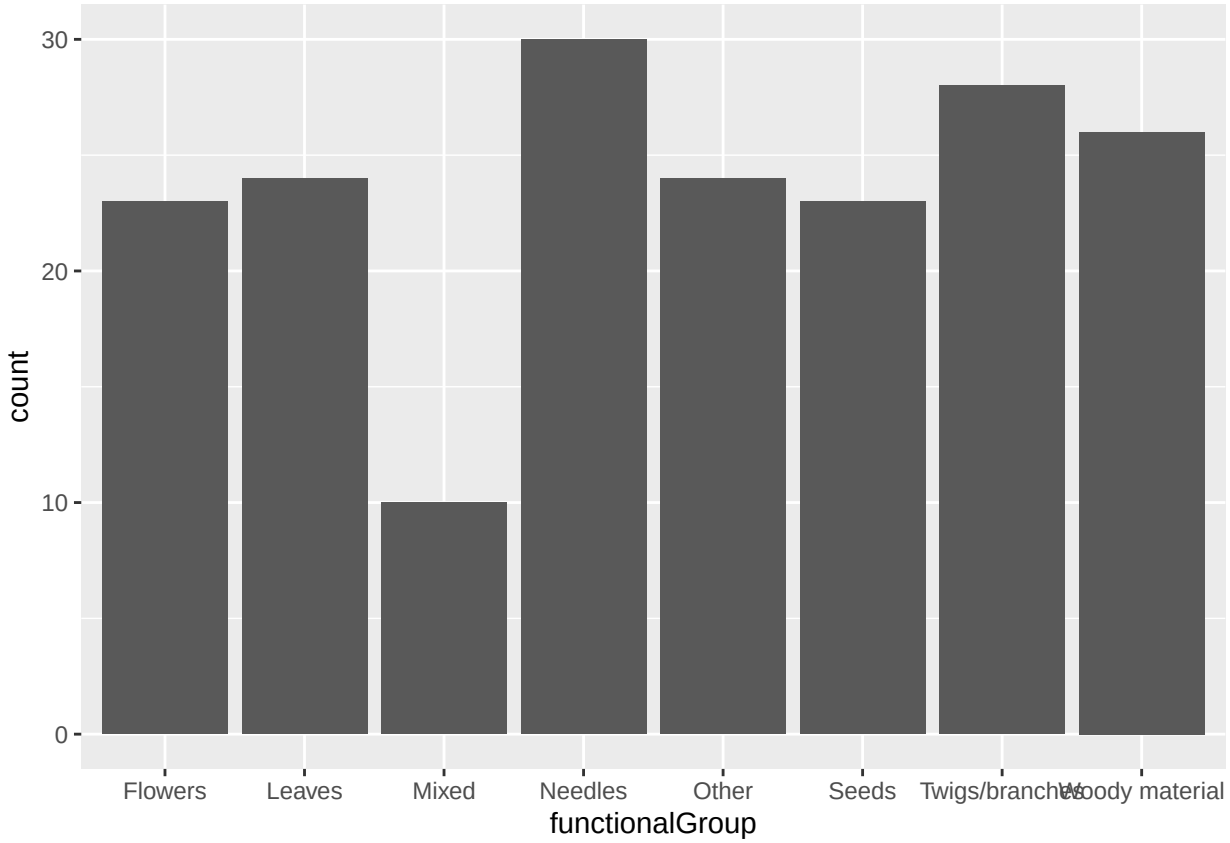
```
length(summary(Litter$plotID)) #Counted number of unique values using length()
```

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

Answer: 12 different plots were sampled at Niwot Ridge. The information obtained from Unique includes the same factors as Summary, but Summary also includes the frequency of each factor in the dataset.

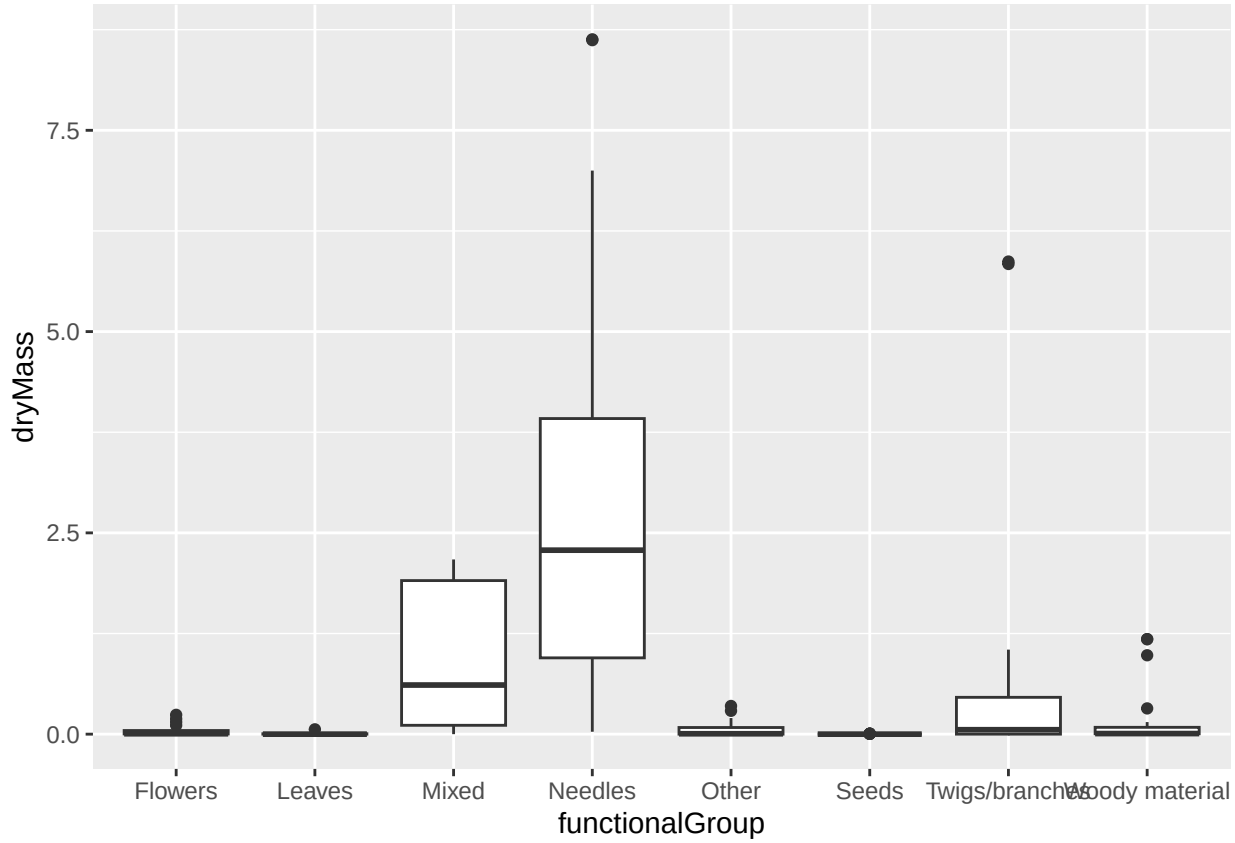
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(x = functionalGroup)) +
 geom_bar() #Counts appear relatively even with exception to "Mixed".
```

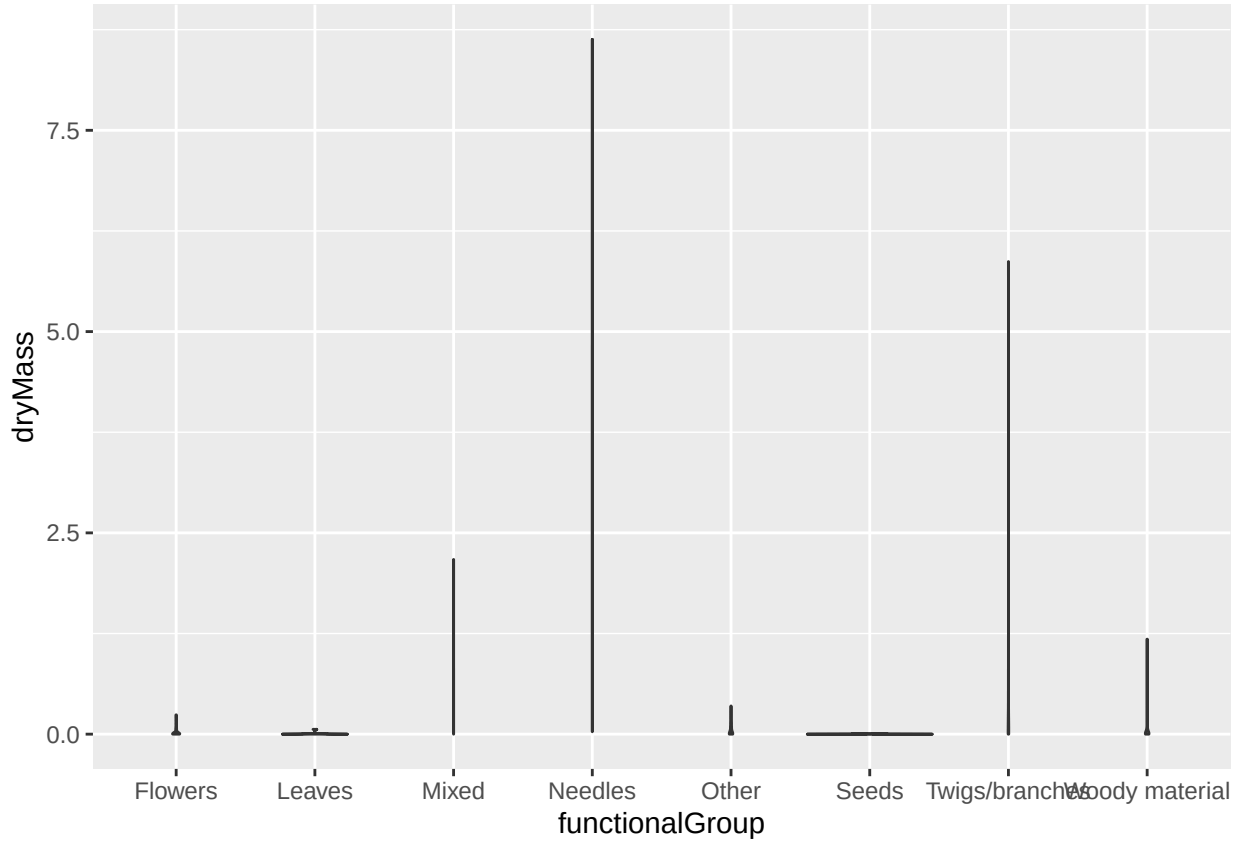


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
 geom_boxplot()
```




```
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
 geom_violin() #Made both graphs and compared shapes
```



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

> Answer: The boxplot is a more effective visualization than the violin plot in this case because of no particular dryMass measurement having a high density/frequency.

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

> Answer: It appears that needles tend to have the highest biomass, with mixed coming in second (based on averages).