Assignment 3: Data Exploration

Ellie harrigan

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.

```
#Load packages and set up working directory

library(tidyverse)
library(lubridate)
library(here)

getwd()
```

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

```r
here()
```

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

```r
Neonics_data <- read.csv(
  file = here('Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv'),
  stringsAsFactors = TRUE)

litter_data <- read.csv(
  file = here('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: One might be interested in ecotoxicology of neonicotinoids on insects to study a variety of factors. Such as, how effective are these at terminating specific insects? Are these poisinous to other species that rely on insects as part of their diets? Do neonicotinoids cause unintended mortality to "beneficial" insects, not just the selective pests?

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 that falls to the ground in forests to study the nutrient content of soil which in turn can help answer questions related to biodiversity within the forest floor. Litter and woody debris can also help answer questions of a forests carbon storage capability, as well as the overall health of the forest ecosystem.

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. Sites with forested tower airtraps sampling took place in 20 40x40 tower plots. Sampling at sites with low-statured vegetation over the tower airsheds were targeted to take place in 4 40x40m tower plots plus 26 20x20m plots.

   2. Once every five years at a site in October or during peak senescence (usually occurrs in the Fall) one round of sampling took place where litter and woody debris was selected for additional processing and analysis by external labs. Material from one elevated trap and 2 functional groups (leaves/needles) per plot were sent for chemical analyses. This resulted in no more than 60 samples/site/5 years.

   3. Ground traps were sampled 1x/year.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics_data)
```

```
## [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? [Tip: The `sort()` command is useful for listing the values in order of magnitude...]

```
sort(summary(Neonics_data$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
```

```
# This command displays the reactions that are studied in the "Effect" column,
# the values tell us how frequently this effect was recorded.
```

Answer: The most common effects that are studied are mortality and population. These effects could be of specific interest because they can help give an understanding to the health of an insect population and if the neonicotinoids are unintentionally harming "beneficial" 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...]

```
sort(summary(Neonics_data$Species.Common.Name), 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
```

```
# Displays the column "species by common name" in order of magnitude,
# decrasing = TRUE shows counts of greatest to least.
```

Answer: The most commonly studied species in the dataset are the Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and the Italian Honeybee. What these species have in common is they are all part of the order Hymenoptera. These insects might be of interest over others because they are valued as “indicator species” and are vital to the health of ecosystems through their roles as pollinators. If these species populations are declining it may indicate that the ecosystem is more vulnerable to outside threats. Bees are also highly susceptible to neonicotinoids due to their foraging behaviors which may present another reason why they are the most popular studied species in the dataset.

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_data$Conc.1..Author.)
```

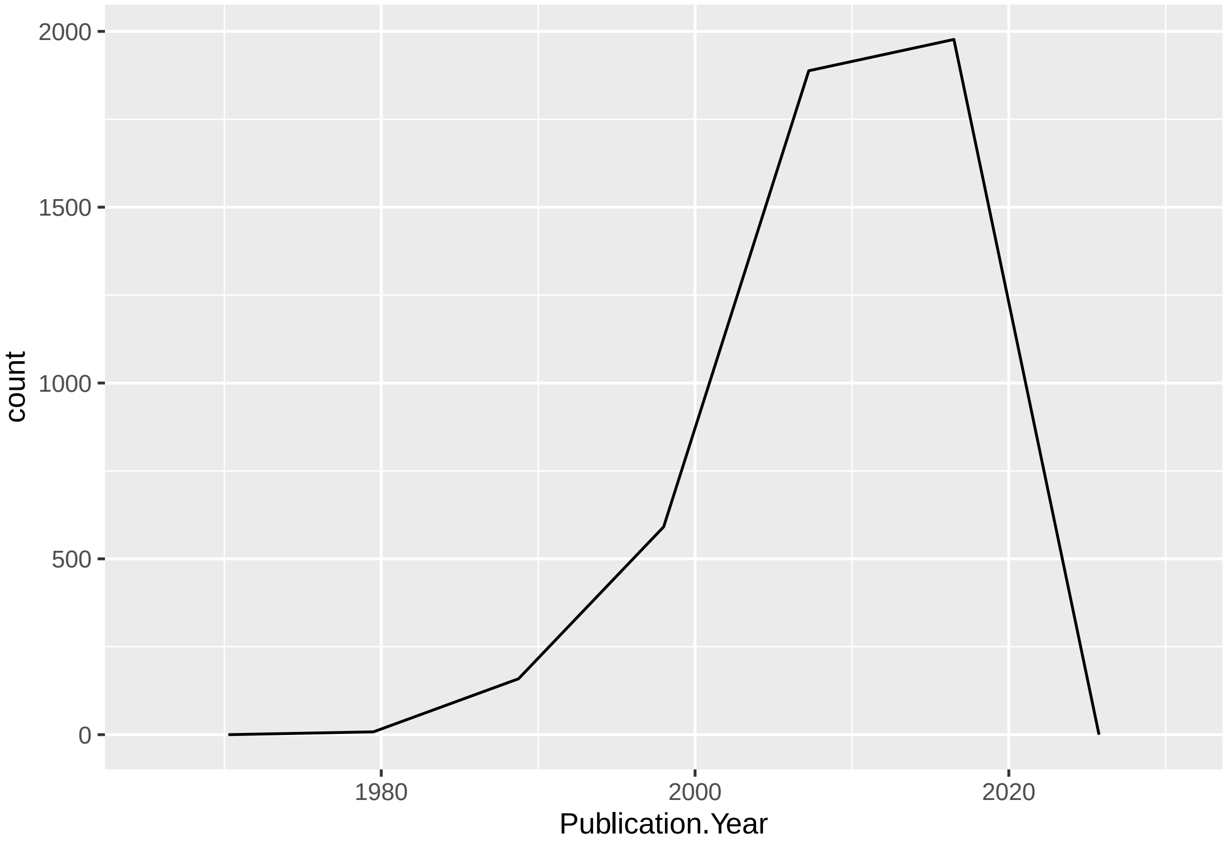
```
## [1] "factor"
```

> Answer: The class is “factor” it is not numeric because of the stringsAsFactors = TRUE command which attaches factors to strings in the dataset. In the Conc.1..Author column there are strings, hence it is assigned as a factor.

Explore your data graphically (Neonics)

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

```
Q9 <- ggplot(Neonics_data) +
  geom_freqpoly(aes(x = Publication.Year), bins = 5)

print(Q9) #this function will make sure the graphs are displayed when I knit the document.
```



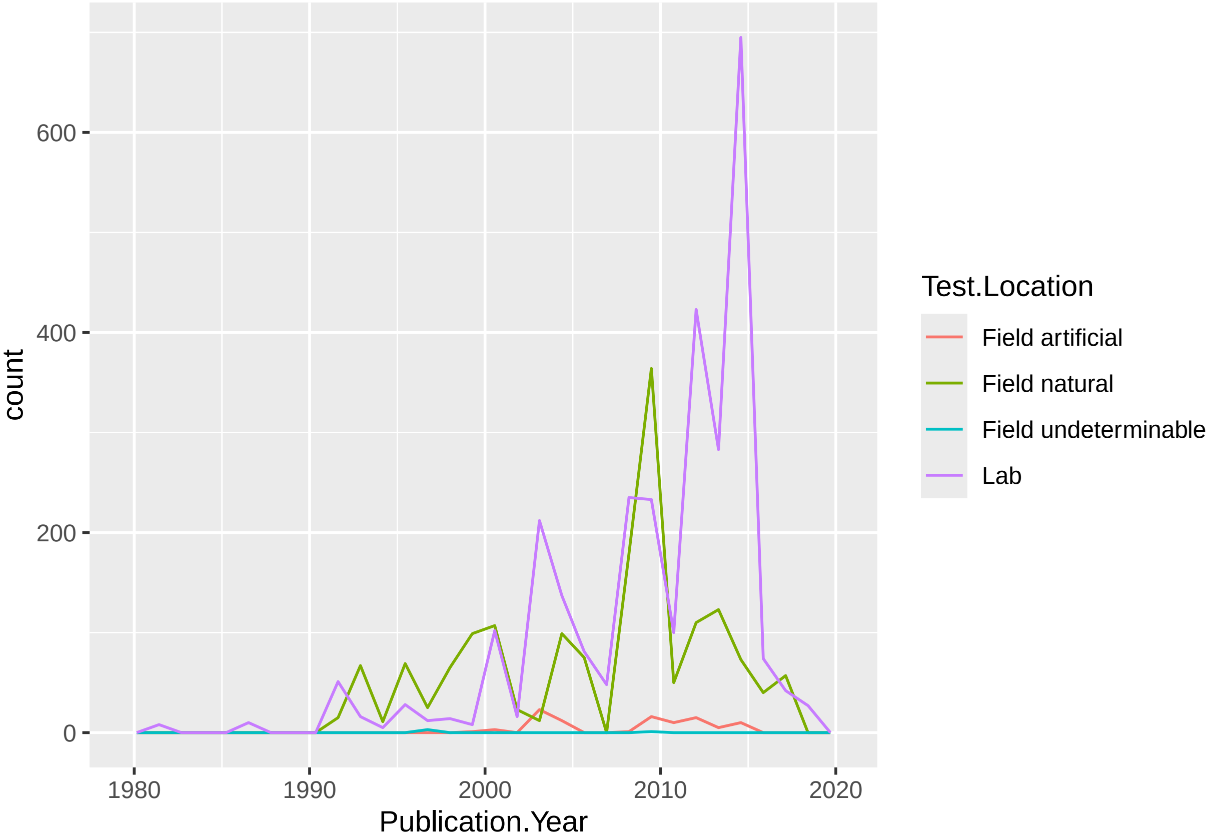
10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
Q10 <- ggplot(Neonics_data) +
  geom_freqpoly(aes(x = Publication.Year, bins = 5, color = Test.Location))
```

```
## Warning in geom_freqpoly(aes(x = Publication.Year, bins = 5, color =
## Test.Location)): Ignoring unknown aesthetics: bins
```

```
print(Q10)
```

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



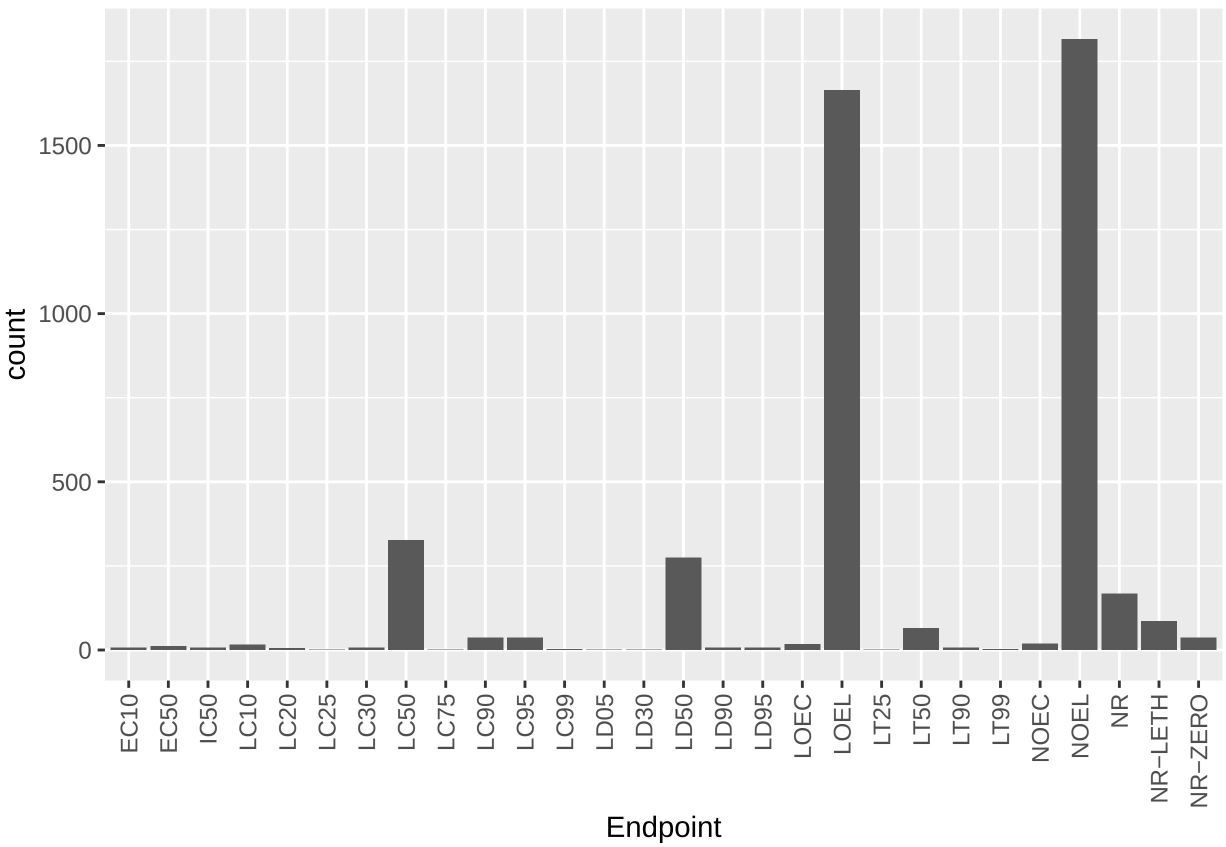
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are "Lab" and "Field Natural". They follow a similar declining trend overtime, specifically after 2015, however "Field Natural" has a sharper decline after 2010 while "Lab" has a very sharp decline after 2015.

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...]

```
Q11 <- ggplot(data = Neonics_data, aes(x = Endpoint)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))

print(Q11)
```



Answer: The most common end points are LOEL and NOEL. LOEL is defined as the lowest-observable-effect-level, meaning the lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC). NOEL is defined as No-observable-effect-level, meaning highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test (NOEAL/NOEC) source: ECOTOX_CodeAppendix, 2019, pp. 722-723.

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_data$collectDate)
```

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

```
litter_data$collectDate <- as.Date(litter_data$collectDate, format('%Y-%m-%d'))
class(litter_data$collectDate)
```

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

```
# This function assigns the litter_data$collectDate the reformatted values for Dates

unique(litter_data$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_data$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
```

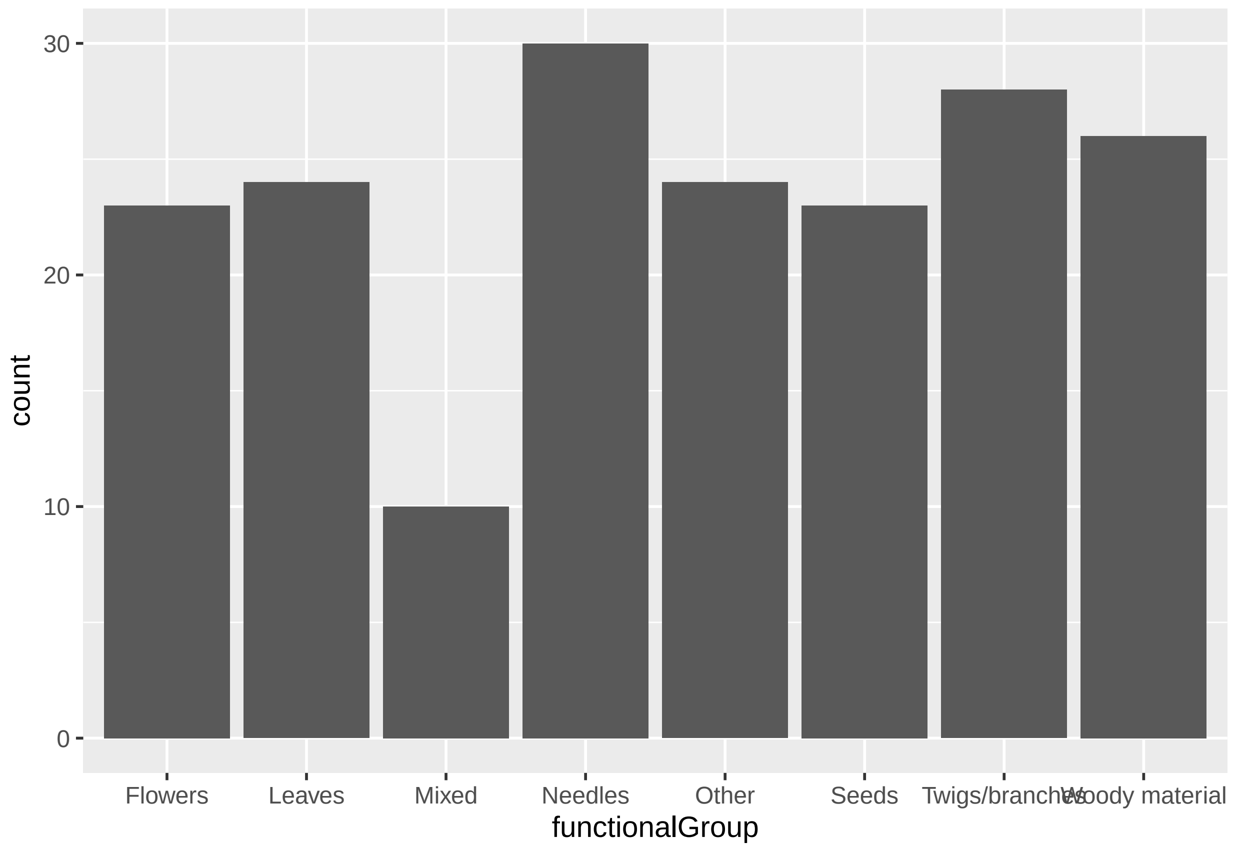
```
summary(litter_data$plotID)
```

```
## 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 different plots. This information obtained using the 'unique' function is different from the summary function in that the unique function displays the different values in the column without repeating them, and shows you potential values that are not present but may be. The summary function shows the exact values in the column and the number of times they are repeated. The summary function is more useful here to determine how many different plots were sampled at Niwot Ridge.

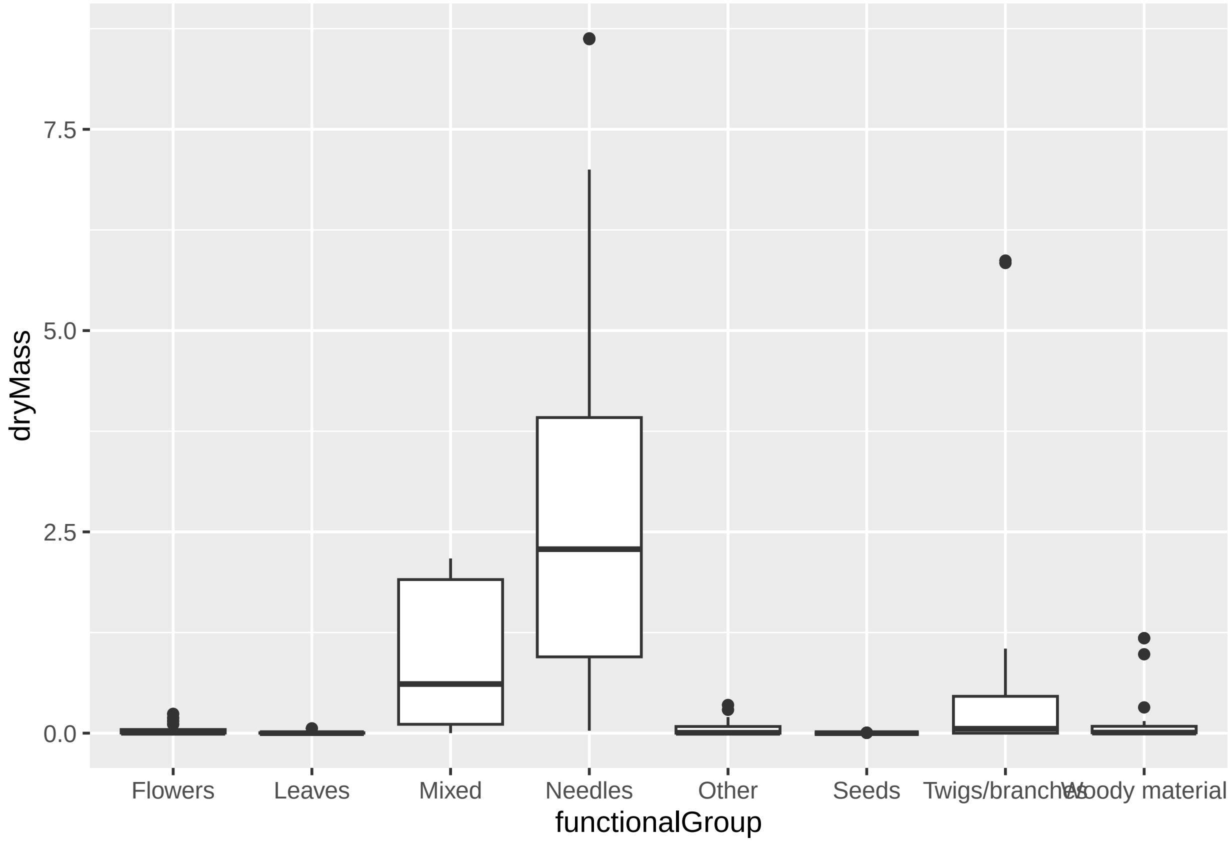
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
Q14 <- ggplot(litter_data, aes(x = functionalGroup)) +
  geom_bar()

print(Q14)
```

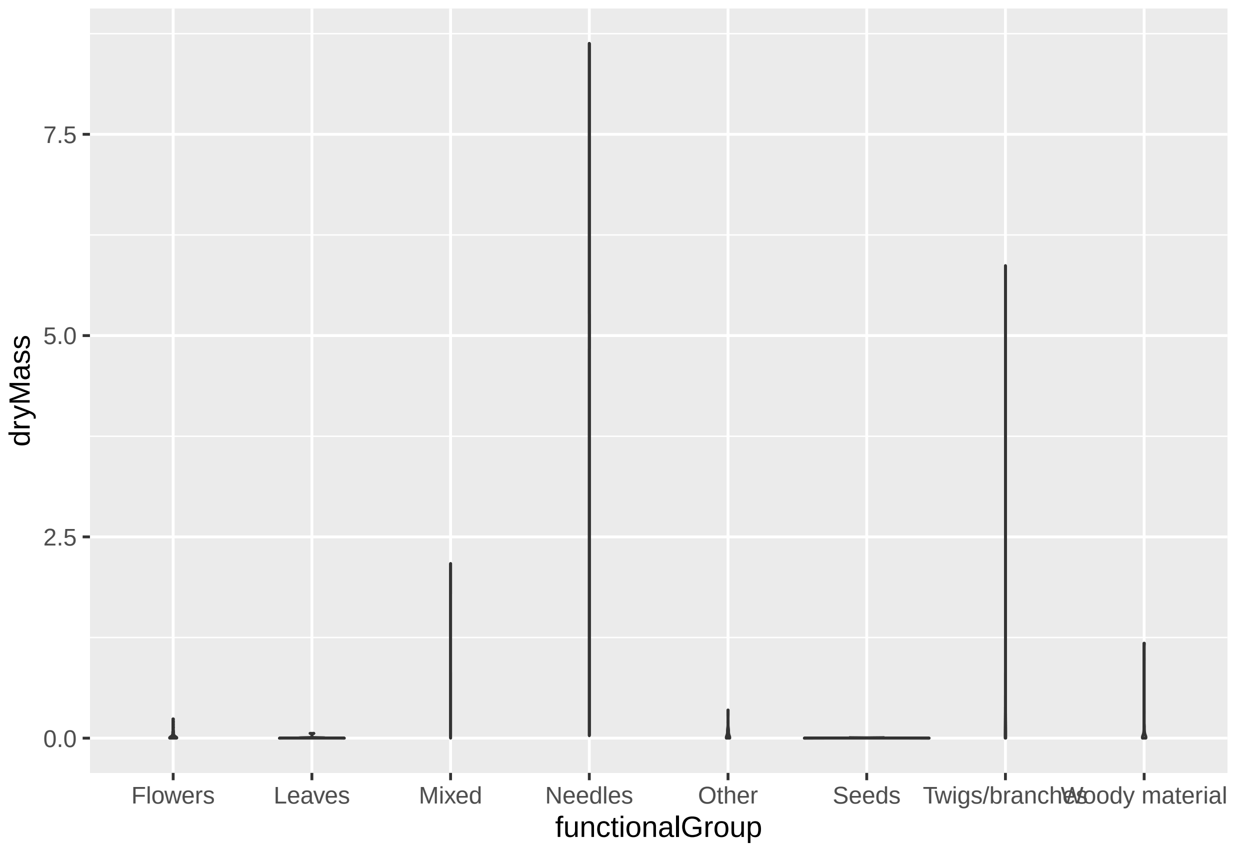


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
Q15_A <- ggplot(litter_data) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass))

print(Q15_A)
```



```
#this function is useful because I have two graphs in the same data chunk,
# when I go to knit the document only the second graph shows
# until I gave the plots a named variable.

Q15_B <- ggplot(litter_data) +
  geom_violin(aes(x = functionalGroup, y = dryMass))

print(Q15_B)
```



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

> Answer: The boxplot is a more effective visualization because it displays the summary statitics in a more informative way. The violin plot displays density values, which in this case appear to be thin lines on the graph, not displaying information on the mean value, outliers of the data, or where the IQR lie. Thus, the boxplot gives us a better understanding of summary statistics for the litter biomass.

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

> Answer: Needles and mixed vegetation have the highest biomass at these sites.