Assignment 3: Data Exploration

Samantha Jensen

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.

```
# Loading needed packages for assignment and checking current working directory with getwd()
library(tidyverse)
```

```
## -- 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()
```

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

```
#Uploading 2 datasets
Neonics <- read.csv(
  file = here("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"),
  stringsAsFactors = TRUE)

Litter <- read.csv(
  file = here("./Data/Raw/NEON_NIWO_Litter_massdata_2018_08_raw.csv"),
  stringsAsFactors = TRUE)
#view(Neonics)
#view(Litter)
```

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: The ecotoxicology of neoninotinoids is important to understand their impact on different species of insects in order to know if they would be useful to implement to combat certain species. Additionally, it is important to understand thier toxicology in order to mitigate the risk of eliminating important and non harmful insect species.

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: Studying the litter and woody debris that falls on the ground in the forests helps us study the link between the tree canopy and the ground soil. The debris falling on the ground is a crucial to the pathway of returning nutrients to the soil and keeping it moist.

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. Plot edges are separated by at least 150% of the plots length 2. Trap placement can be targeted or ransomized depending on the vegetation within the plot. 3. Different types of traps have different sampling frequencies (ex. ground = 1x/yr)

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset? The Dataset Neonics has 4623 rows and 30 columns.

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

```
# Sorting the summary of the effects column
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
```

```
view(Neonics)
```

Answer: The most commons effects are population and mortality. I think that these effects are specifically of interest because the scientists are interested in how individual and groups of insects are impacted by these neoninotinoids. The population effect of the neoninotinoids is measured typically through abundance of insects, while the mortality looks at the mortality metric. The abundance and mortality are insightful determining the impact of neoninotinoids.

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...]

```
#sorting the summary of the species.common.name column
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: All of these species are species are either bees or wasps. I think that they might be of interest over other insects because they have a negative connotation with humans because they sting. While other insects might be seen as "annoying" or "gross," they don't pose a threat to humans.

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...]

```
# checking the class of the Conc.1..Author column
class(Neonics$Conc.1..Author.)
```

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

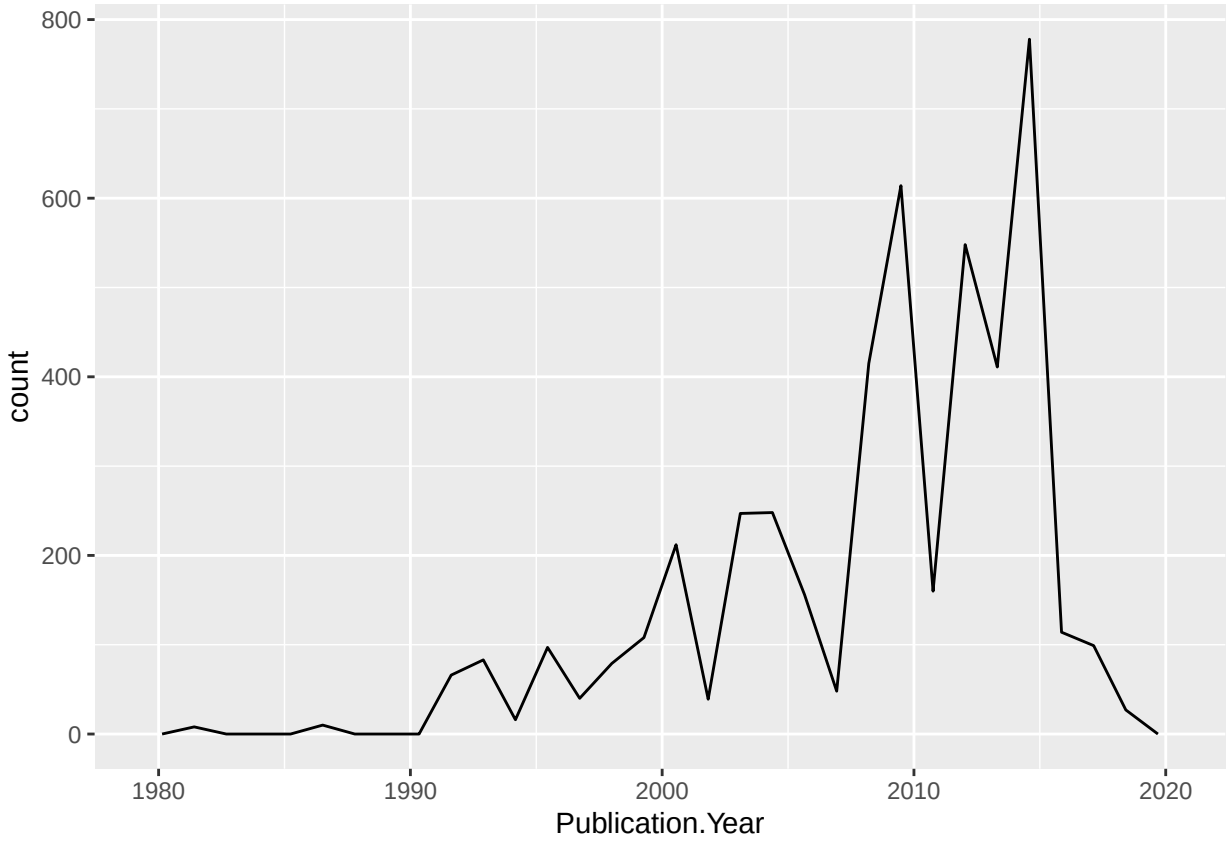
Answer: The class is a Factor. The Conc.1..Author column is not numeric because the values are meant to be represented as categorical data, not integers. No math should ever be done on the column of concentrations because they are experiment specific and not measured values.

Explore your data graphically (Neonics)

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

```
# creating a frequency plot for the count of studies published in each year
ggplot(Neonics, aes(x = Publication.Year)) +
  geom_freqpoly()
```

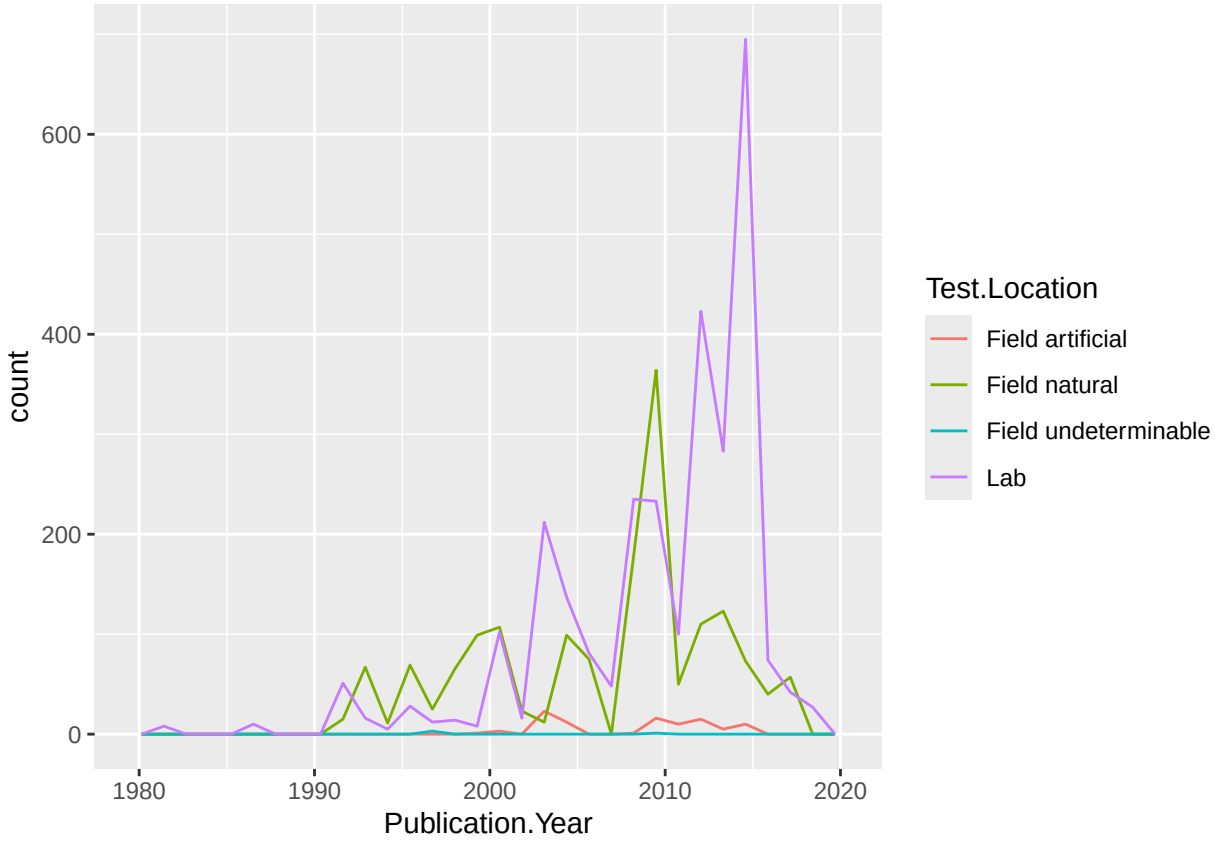
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
# separating counts by test location
ggplot(Neonics, aes(x = Publication.Year, color = Test.Location)) +
  geom_freqpoly()
```

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



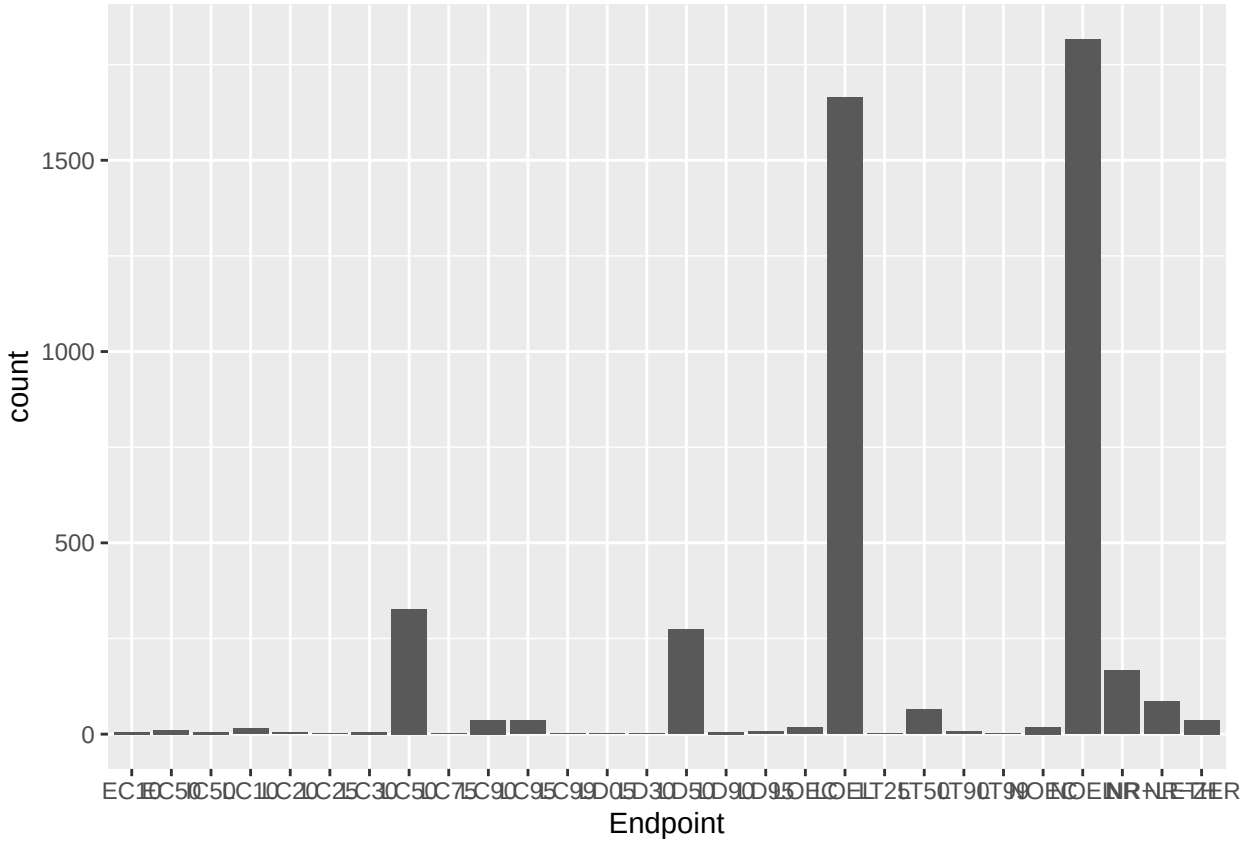
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are in the Lab and naturally in the field. Until ~2010, the count of studies occuring in the lab and naturally in the field were pretty similar, with variations in which one was more year to year. However, after 2010, lab tests have skyrocketed and are the dominant test 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...]

```
# creation of a bar graph based on end points
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar()
```




```
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

```
## List of 1
##  $ axis.text.x:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : num 1
##   ..$ vjust        : num 0.5
##   ..$ angle        : num 90
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

Answer: The most common end points are LOEL and NOEL. LOEL is defined as the Lowest Observable Effect Level which indicates the lowest concentration level that produced an effect that was different from the control by a significant level. NOEL stands for No Observable Effect Level which represents the highest dose level that produces an effect that is not significantly different from the control.

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.

```
 # Checking the class of collectDate Column
 class(Litter$collectDate)
```

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

```
# changing collectDate from factor to date in ymd format
Litter$collectDate <- ymd(Litter$collectDate)
# checking that collectDate is now a date class.
class(Litter$collectDate)
```

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

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`?

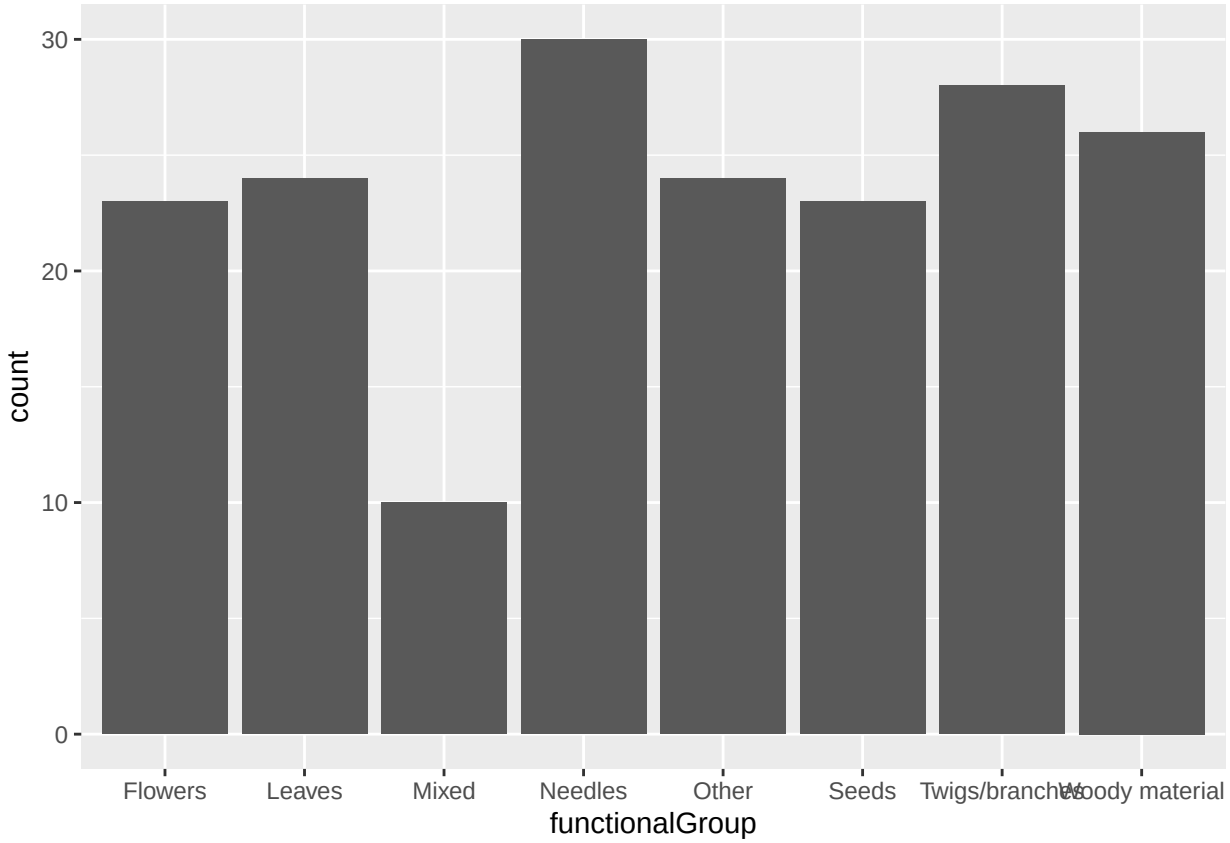
```
# unique() function used to determine how plots were sampled. Found how many unique plot ID codes there
unique(Litter$plotID)
```

```
##  [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
```

> Answer: The unique function produces each individual "value" that exists in the column of interest without any duplicates. The summary function also lists these values and in addition it produces the count of how many times each of the individual plot codes appeared in the dataset.

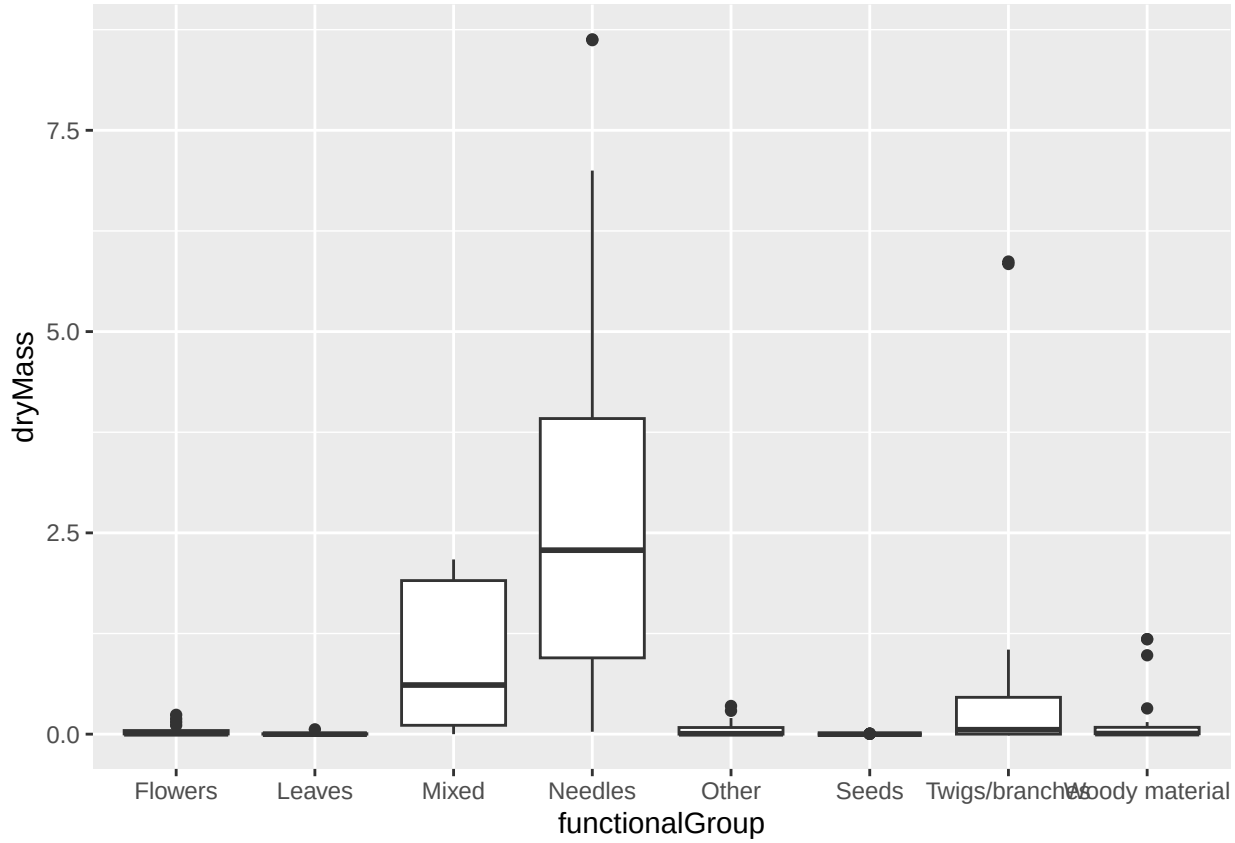
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
# creating a bar plot showing the counts of each functionGroup category.
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar()
```

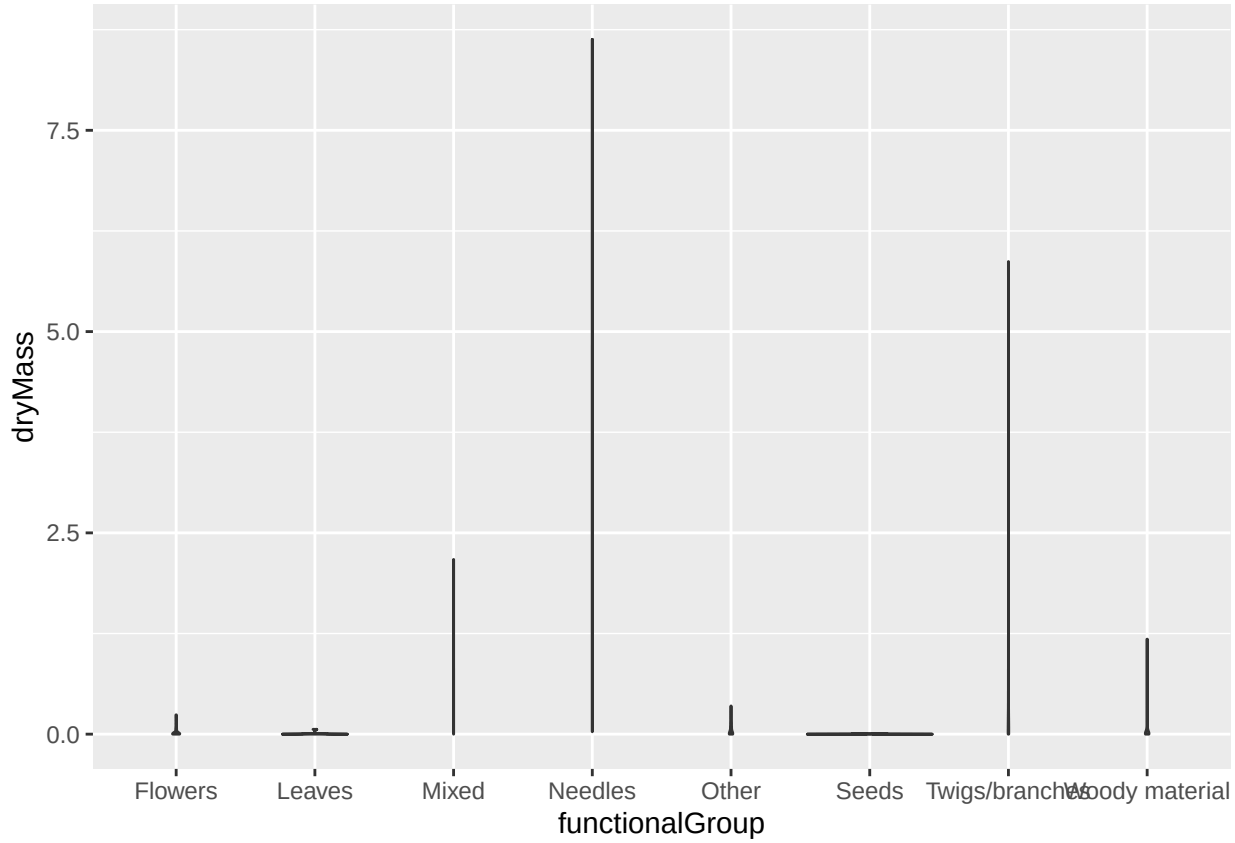


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
# creating a box plot of dry mass in each functional group.
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_boxplot()
```




```
# creating a violin plot of dry mass in each functional group
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_violin()
```



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

> Answer: In this case, the density of the data is not necessary for effectively visualizing the data. The violin plot shows the distribution of data but in this case, the distribution makes the graph almost impossible to read as the 'violins' are unreadable as they take up such a small area on the graph. The box plot effectively presents the summary statistics as well as showing outliers for each group.

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

> Answer: Needles have the highest biomass at these sights with mixed litter having the second highest biomass.