Assignment 3: Data Exploration

Brian Mulu Mutua

Fall 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.

```
#Checking working directory
getwd()
```

```
## [1] "C:/Users/bmm100/Documents/EDE_Fall2023"
```

```
#Loading necessary packages
library(tidyverse)
library(lubridate)
```

```
#Importing Neonics dataset
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors
↪  = TRUE)
#Displaying first 3 rows and first 2 columns of the Neonics dataframe to show successful
↪  import
head(Neonics,c(3,2))
```

```
##   CAS.Number                                    Chemical.Name
## 1   58842209 Tetrahydro-2-(nitromethylene)-2H-1,3-thiazine
## 2   58842209 Tetrahydro-2-(nitromethylene)-2H-1,3-thiazine
## 3   58842209 Tetrahydro-2-(nitromethylene)-2H-1,3-thiazine
```

```
#Importing Litter dataset
Litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv",
↪  stringsAsFactors = TRUE)
#Displaying first 3 rows and first 2 columns of Litter dataframe to show successful
↪  import
head(Litter,c(3,2))
```

```
##                                   uid         namedLocation
## 1 7f065fec-bcb2-4af9-b742-8e520fab7f6e NIWO_061.basePlot.ltr
## 2 88df210b-1445-4c3f-b19e-5dabd9305c6e NIWO_061.basePlot.ltr
## 3 7f3c549c-1dfa-43bf-a485-c7c2bcb31fd6 NIWO_061.basePlot.ltr
```

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: Studies have shown that agricultural use of neonicotinoids as insecticides can have unintended deadly effects on pollinators such as bees or other invertebrate species that play a vital role in ecosystems. Due to the widespread use of neonicotinoids, which are the single most popular insecticide class in the US, it is important to understand how their use impacts ecosystems. Investigating the ecotoxicology of neonicotinoids on insects helps us assess the unexpected target effects of the insecticides on various insect species and resulting impacts to biodiversity and food security.

References *Lindwall C. (2022, May 25). Neonicotinoids 101: The Effects on Humans and Bees. NRDC. https://www.nrdc.org/stories/neonicotinoids-101-effects-humans-and-bees*

Francisco S., Gaka K., Hayasaka D., (2016, Nov 02). Contamination of the Aquatic Environment with Neonicotinoids and its Implication for Ecosystems. Frontiers in Environmental Science. https://www.frontiersin.org/articles/10.3389/fenvs.2016.00071/full

Frank S. D., Tooker J. F., Neonicotinoids pose undocumented threats to food webs. PNAS. https://www.pnas.org/doi/full/10.1073/pnas.2017221117

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: Litter refers to the amount of leaves, buds, twigs, seeds, and flowers that fall from trees. In forest ecosystems, litter is an important component of the nutrient cycle that regulates the accumulation of soil organic matter (SOM), the input and output of the nutrients, nutrient replenishment, biodiversity conservation and other ecosystem functions. Litter serves as a crucial input of carbon and nutrients to the forest floor. Therefore, an understanding of the major processes (litterfall production and its decomposition rate) in the cycle is vital for sustainable forest management (SFM).

References *Giweta M. (2020, May 07). Role of litter production and its decomposition, and factors affecting the processes in a tropical forest ecosystem: a review. BMC. https://jecoenv.biomedcentral.com/articles/10.1186/s41610-020-0151-2*

Zukswert J. (2023, April 26). LITTERFALL IN THE LIMELIGHT: HOW A "COVID PAPER" FROM HARVARD FOREST SHEDS LIGHT ON SPATIAL AND TEMPORAL LITTERFALL PATTERNS. LTER NETWORK. https://lternet.edu/stories/litterfall-in-the-limelight-how-a-covid-paper-from-harvard-forest-sheds-light-on-spatial-and-temporal-litterfall-patterns/

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.Samples are used to provide mass data for plant functional groups from individual sampling bouts, measured to an accuracy of 0.01 grams. Weights < 0.01g are reported to indicate presence of a functional group, but not at detectable masses. Litter and fine woody debris are collected from elevated and ground traps, respectively. Litter is defined as material that is dropped from the forest canopy and has a butt end diameter <2cm and a length <50cm collected in elevated 0.5m2 PVC traps. Fine wood debris is defined as material that is dropped from the forest canopy and has a butt end diameter <2cm and a length >50cm collected in ground traps.
   2.Sampling for litter and fine woody debris occurs only in tower plots that are *selected randomly* within the 90% flux footprint of the primary and secondary airsheds (and additional areas in close proximity to the airshed, as necessary to accommodate sufficient spacing between plots.)
   3.Ground traps are sampled once per year. Target sampling frequency for elevated traps varies by vegetation present at the site, with frequent sampling (once every 2 weeks) in deciduous forest sites during senescence and in-frequent year-round sampling (once every 1-2 months) at evergreen sites.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#Checking and printing out the dimensions ofthe Neonics dataset
print("Dimensions of the Neonics dataset are: ")
```

```
## [1] "Dimensions of the Neonics dataset are: "
```

```r
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?

```r
sort(summary(Neonics$Effect),decreasing = TRUE)
```

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

Answer: The two most common effects studied include Population and Mortality at 1803 and 1493 studies respectively. These are followed by Behavior, Feeding behavoir and Reproduction which wrap up the top 5 effects studied at 360, 255 and 197 studies respectively. Population and Mortality would specifically be of interest as they are the key indicators to establish impacts of neonicotinoids on insect population and mortality. Behavoir, Feeding behavior and Reproduction effects can provide insights into additional adverse impacts that are not immediately fatal.

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...]

```r
#Using summary function to determine the six most commonly studied species
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
```

Answer: The six most studied species by common name as shown include the Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee and Italian Honeybee. The top six species are either bees or wasps that belong in the same order of insects and act as pollinators that support food systems. Hence, such pollinator species would be of particular interest over other insects when assessing impacts of the use of neonicotinoids.

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?

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

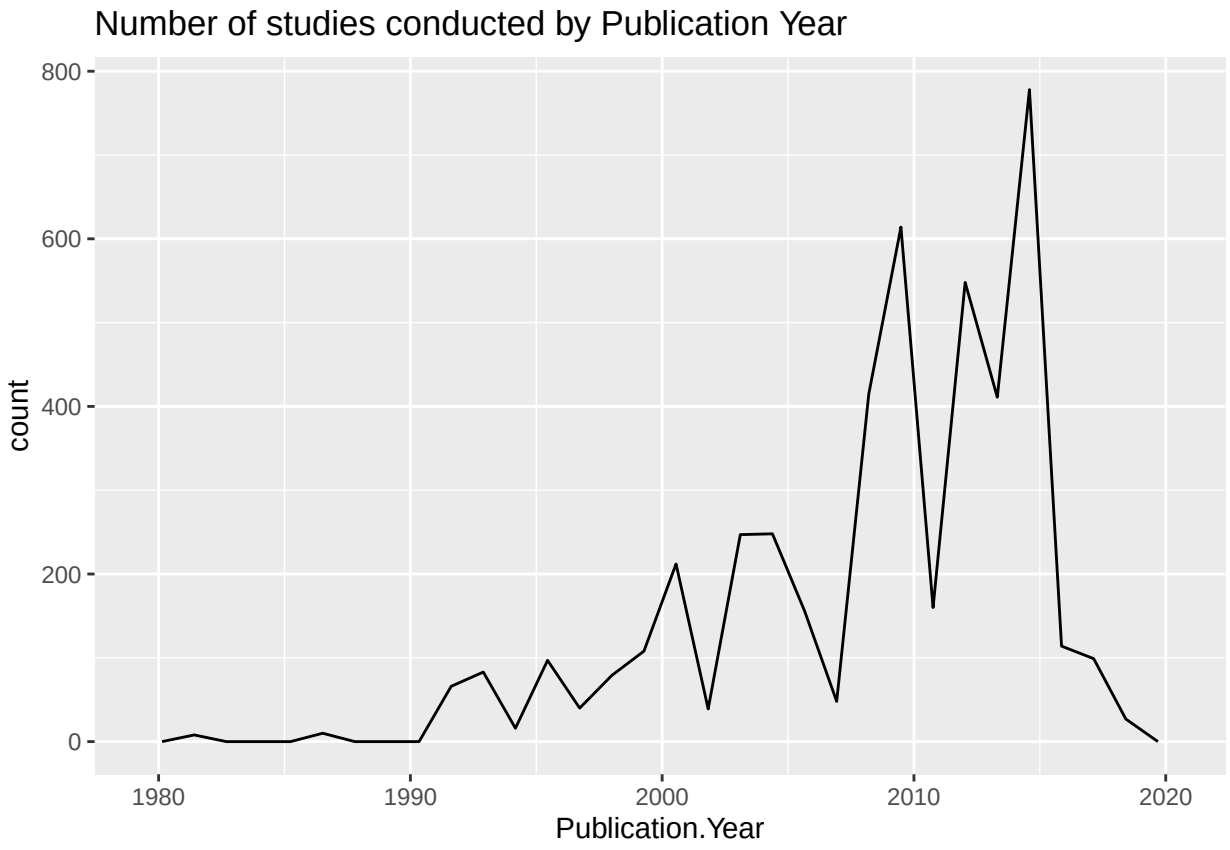
```
## [1] "factor"
```

Answer: The class of the `Conc.1..Author.` column in the dataset is **factor**. The reason it is not stored as a numeric value is due to the presence of non-numeric characters in the dataset such as "/" or "<", which led R to import the values as factors instead of numeric values.

Explore your data graphically (Neonics)

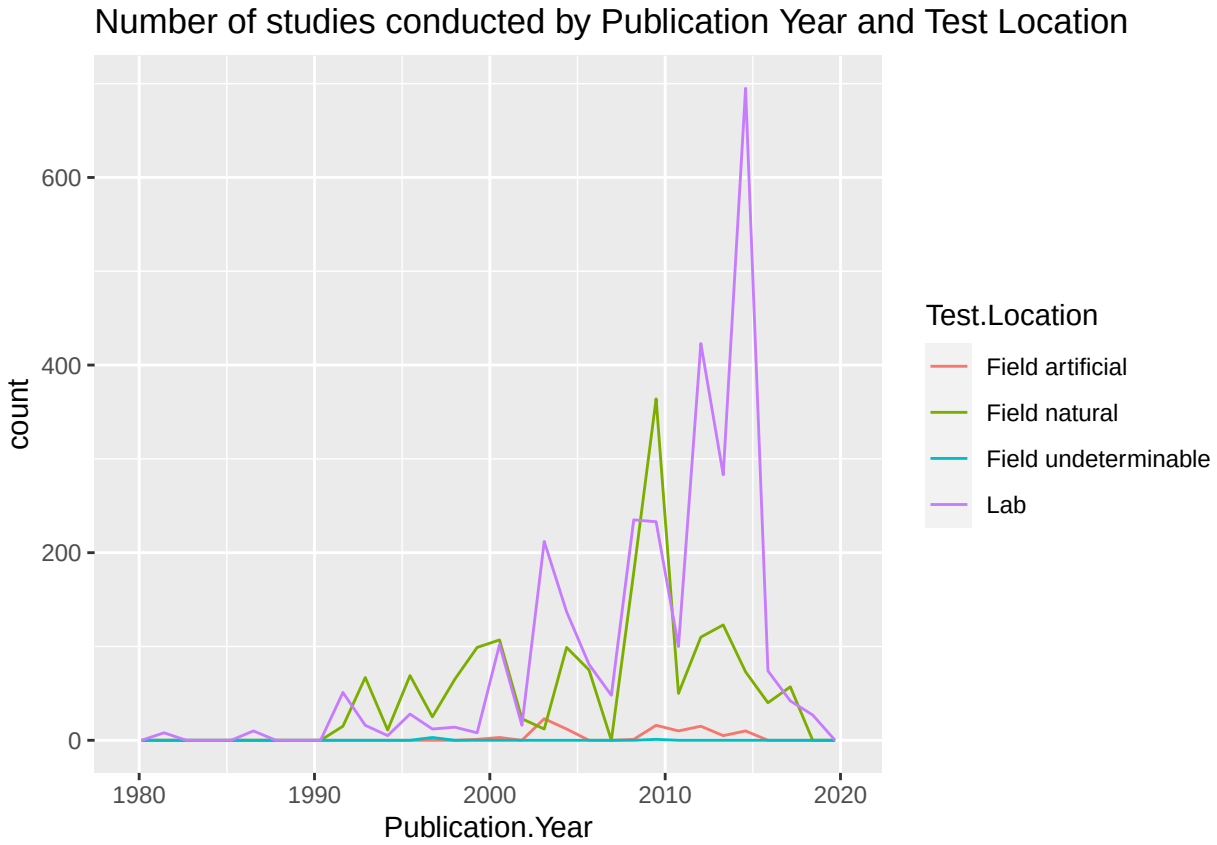
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```r
ggplot(Neonics) + geom_freqpoly(aes(x=Publication.Year), bins = 30) + labs(title =
↪  "Number of studies conducted by Publication Year")
```



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

```r
ggplot(Neonics) + geom_freqpoly(aes(x=Publication.Year, color = Test.Location), bins =
↪  30)+ labs(title = "Number of studies conducted by Publication Year and Test
↪  Location")
```



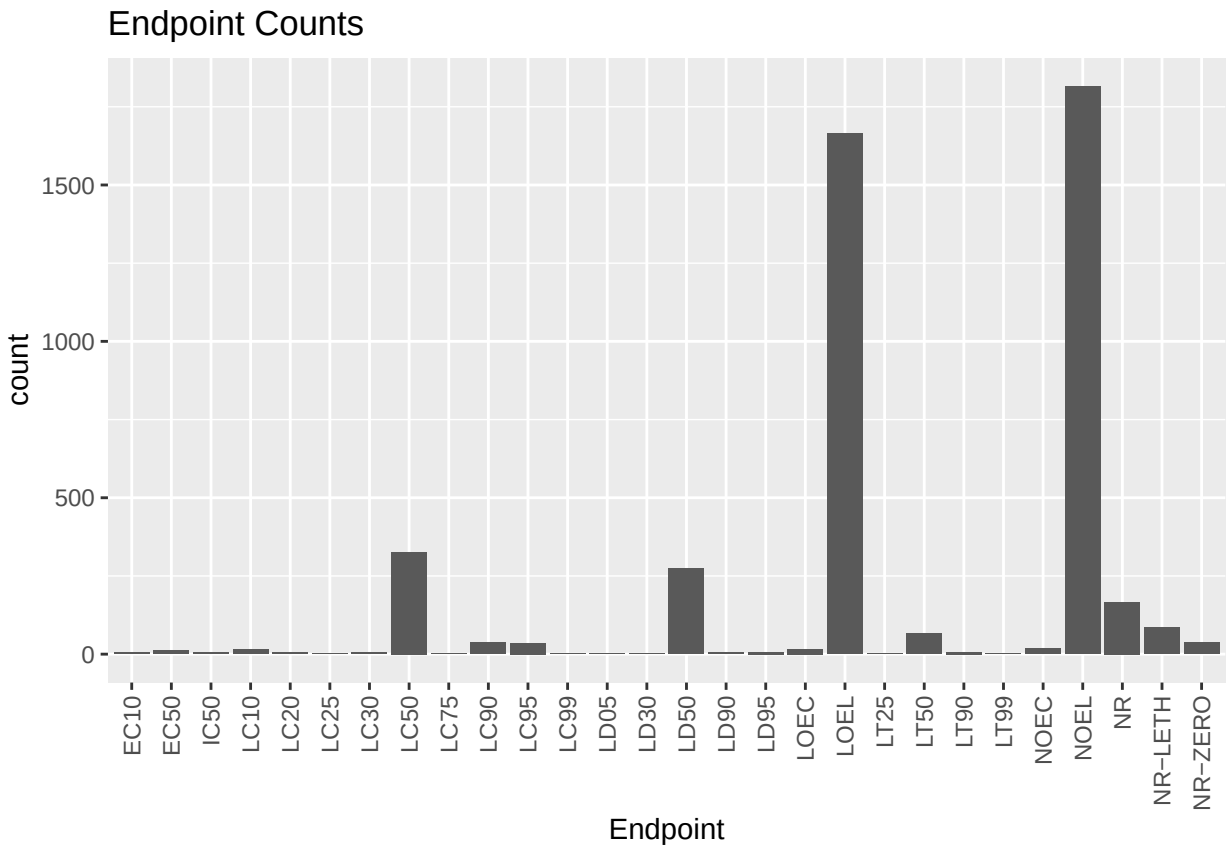
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are *Lab* and *Natural Field*. Test locations of *Artificial Field* and *Undeterminable Field* feature less prominently. In the mid-late 90's, more tests were more commonly conducted in the Natural Field than the other test locations. With time, a higher number of tests have been conducted in the Lab than the Natural Field, particularly in the last decade.

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) + geom_bar(aes(x=Endpoint)) + theme(axis.text.x = element_text(angle =
↪  90, vjust = 0.5, hjust=1)) + labs(title = "Endpoint Counts")
```



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

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
class(Litter$collectDate)
```

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

```
#Printing out a sample value to check format before changing class
print(Litter$collectDate[1])
```

```
## [1] 2018-08-02
## Levels: 2018-08-02 2018-08-30
```

```
#Changing the class of collectDate to the Date class using the lubridate package
Litter$collectDate <- ymd(Litter$collectDate)
#Confirming the new class of collectDate as Date
class(Litter$collectDate)
```

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

```
#Determining dates that the litter was sampled in August 2018 using 'unique' function
print("The dates that the litter was sampled in August 2018 include:")
```

```
## [1] "The dates that the litter was sampled in August 2018 include:"
```

```
unique(Litter$collectDate)
```

```
## [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`?

```
#Using the 'unique' function on the plotID variable to determine number of unique plots
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
```

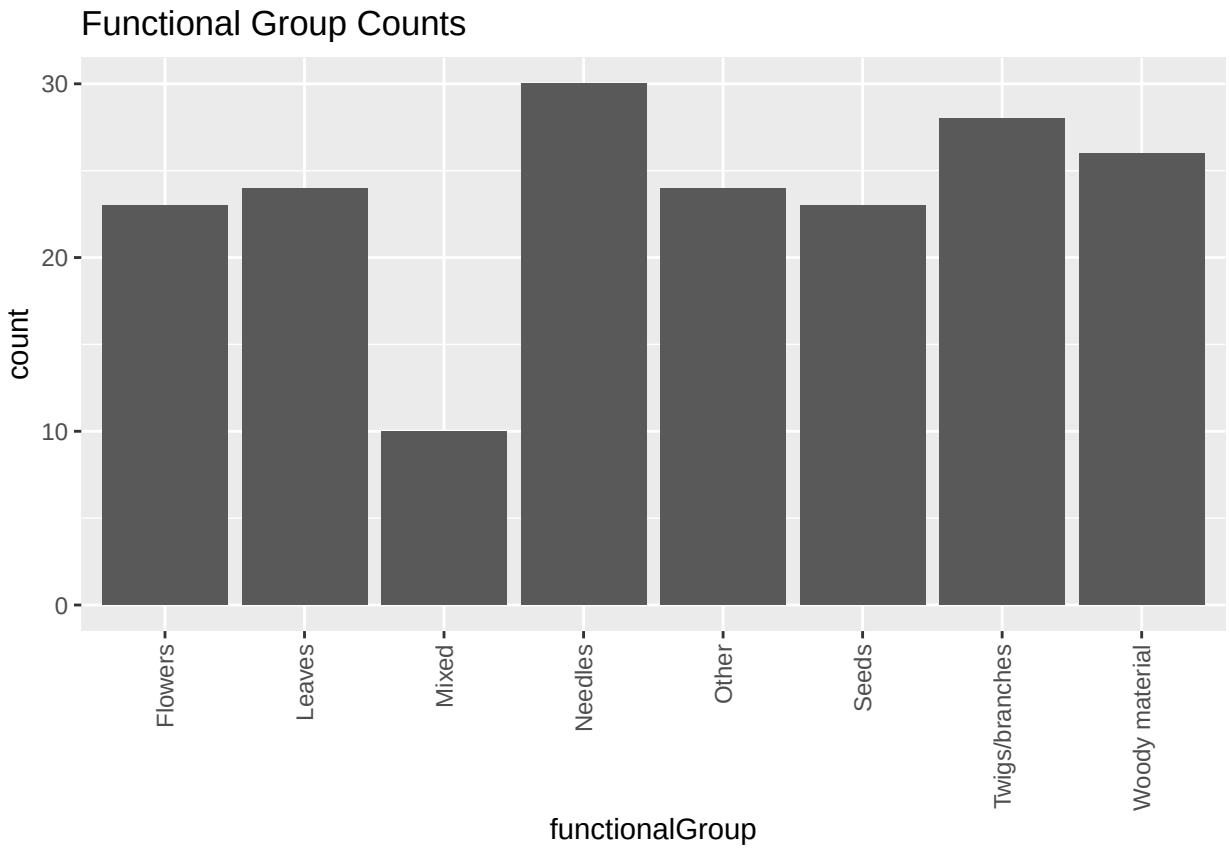
```
#Using the 'summary' function on the plotID variable
summary(Litter$plotID)
```

```
## 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: **12** plots were sampled at Niwot Ridge. The unique function only returns a vector with the unique values and total quantity of the plot IDs which were sampled. This is different from the summary function, which for the same variable, returns not only the plot IDs but also the count of the frequency at which each plot represented by a plotID was sampled.

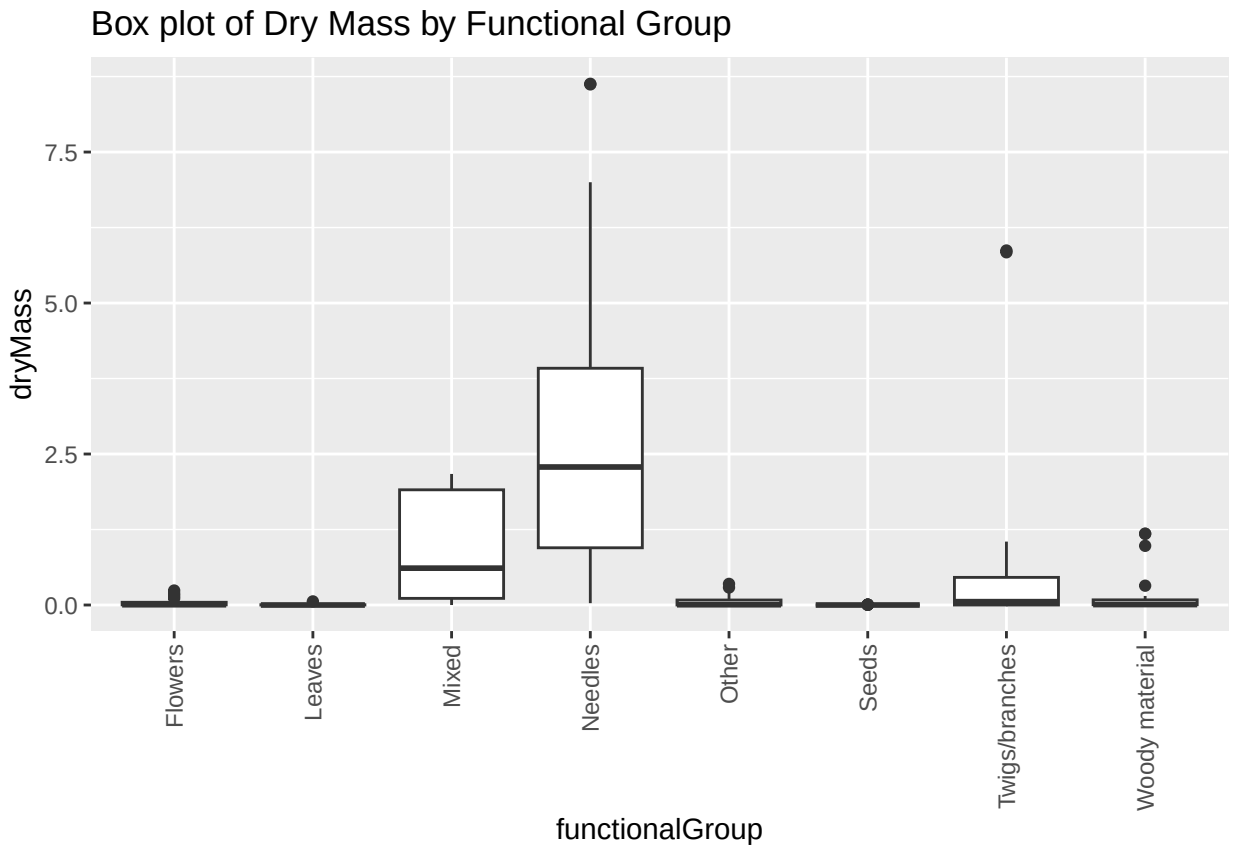
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter) + geom_bar(aes(x=functionalGroup)) + theme(axis.text.x =
↪  element_text(angle = 90, vjust = 0.5, hjust=1)) + labs(title = "Functional Group
↪  Counts")
```

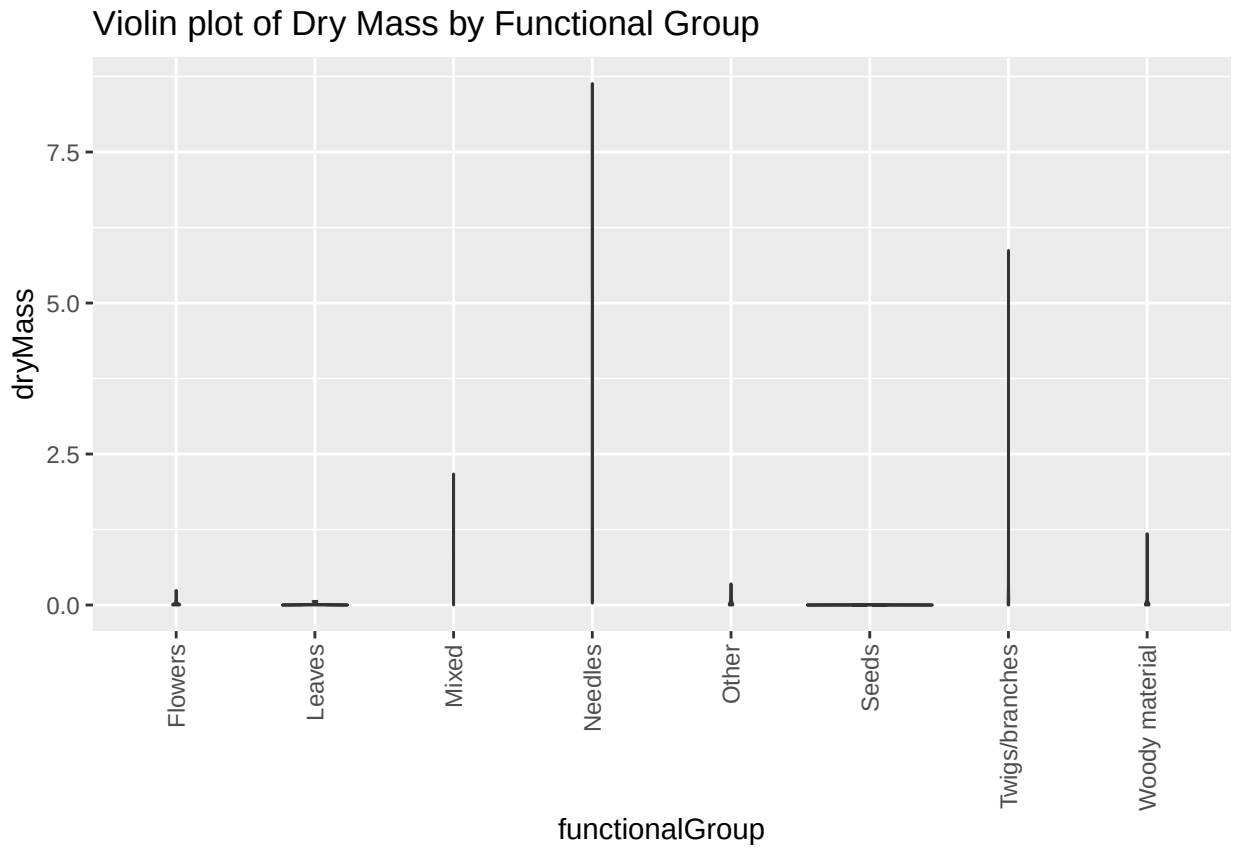


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(Litter) + geom_boxplot(aes(x=functionalGroup,y=dryMass)) + theme(axis.text.x =
↪  element_text(angle = 90, vjust = 0.5, hjust=1)) + labs(title = "Box plot of Dry Mass
↪  by Functional Group")
```




```
ggplot(Litter) + geom_violin(aes(x=functionalGroup,y=dryMass)) + theme(axis.text.x =
↪  element_text(angle = 90, vjust = 0.5, hjust=1)) + labs(title = "Violin plot of Dry
↪  Mass by Functional Group")
```



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

> Answer: The boxplot is a more effective visualization in this case because the functionalGroup values of dryMass are not widely distributed along the range of sampled values. As can be seen from the boxplot above, a lot of the dryMass quantities are densely located closer to 0.0 for most of the functionalGroup categories. As the violin plot works best to illustrate the shape of the distribution of the data, where densities vary over a range,it is less effective as a visualization tool for this particular dataset.

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

> Answer: Based on the visualization, **Needles** tend to have the highest biomass at these sites.