

Assignment 3: Data Exploration

Aditi Jackson

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.

```
# confirming that working directory is my ENV872 folder using getwd()
getwd() # output: "/Users/aditijackson/ENV872 Setup (local)"
```

```
## [1] "/Users/aditijackson/ENV872 Setup (local)"
```

```
# installing packages
# install.packages("dplyr") # helps with data manipulation
# install.packages("ggplot2") # helps with data visualization
# install.packages("tidyverse") # helps with data manipulation and visualization
# install.packages("lubridate") # helps with manipulating date objects
```

```
# loading packages
library("dplyr")
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
```

```
library("ggplot2")
library("tidyverse")
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v forcats   1.0.0      v stringr   1.5.0
## v lubridate 1.9.2      v tibble   3.2.1
## v purrr     1.0.2      v tidyr    1.3.0
## v readr     2.1.4
```

```
## -- 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")
```

```
# Importing data sets
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = TRUE)
# ECOTOX neonicotinoid dataset

Litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv", stringsAsFactors = TRUE)
# Niwot Ridge NEON dataset
```

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: Given that neonicotinoids are widely used in agriculture, they likely have a huge impact on insect populations and ecosystem health more broadly. Given that insects play a critical role in most ecosystems, a decline in or disruption to insect populations could have negative environmental knock-on effects. Studying this data set could help us identify and better understand environmental externalities of using neonicotinoids.

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 may be interested in studying litter and woody debris for a variety of different reasons, including gaining insight into forest biodiversity or conducting fire risk assessments. Overall this data could help inform better forest management practices.

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. Litter and debris sampling is executed at sites that contain woody vegetation greater than 2 meters tall; samples are collected using a combination of ground and elevated traps 2. Sampling only occurs in tower plots which are randomly selected 3. Ground traps are sampled once a year while elevated traps are sampled one time every two weeks

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) # dim() gives use the number of rows and columns in data set
```

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

```
# Neonics has 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?

```
summary(Neonics$Effect) # the "$" allows us to pull a specific column of the Neonics dataset
```

```
##      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
```

Answer: The most commonly studied effects include Population and Mortality. Population is likely of interest to researchers b/c it can help them better understand the effects of neonicotinoids on current and historical insect populations and identify trends. Similarly, researchers are likely interested in studying the effects of neonicotinoids on mortality to identify the effectiveness of insecticides.

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...]

```
Summary_CommonName <- summary(Neonics$Species.Common.Name) # summarizing common names column
sort(Summary_CommonName) # sorting Summary_CommonName, output is in lowest to highest order
```

```
##          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 Woolly 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: The most commonly studied species are the Honey Bee, the Parasitic Wasp, the Buff Tailed Bumblebee, the Carniolan Honey Bee, the Bumble Bee, and the Italian Honeybee. All six belong to the Hymenoptera order of insects (according to Google) and play a crucial role in pollinating agricultural crops. They are likely of higher interest than other insects given they are crucial to the health of plant ecosystems.

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?

```
class(Neonics$Conc.1..Author.) # class of Conc.1..Author. is a factor
```

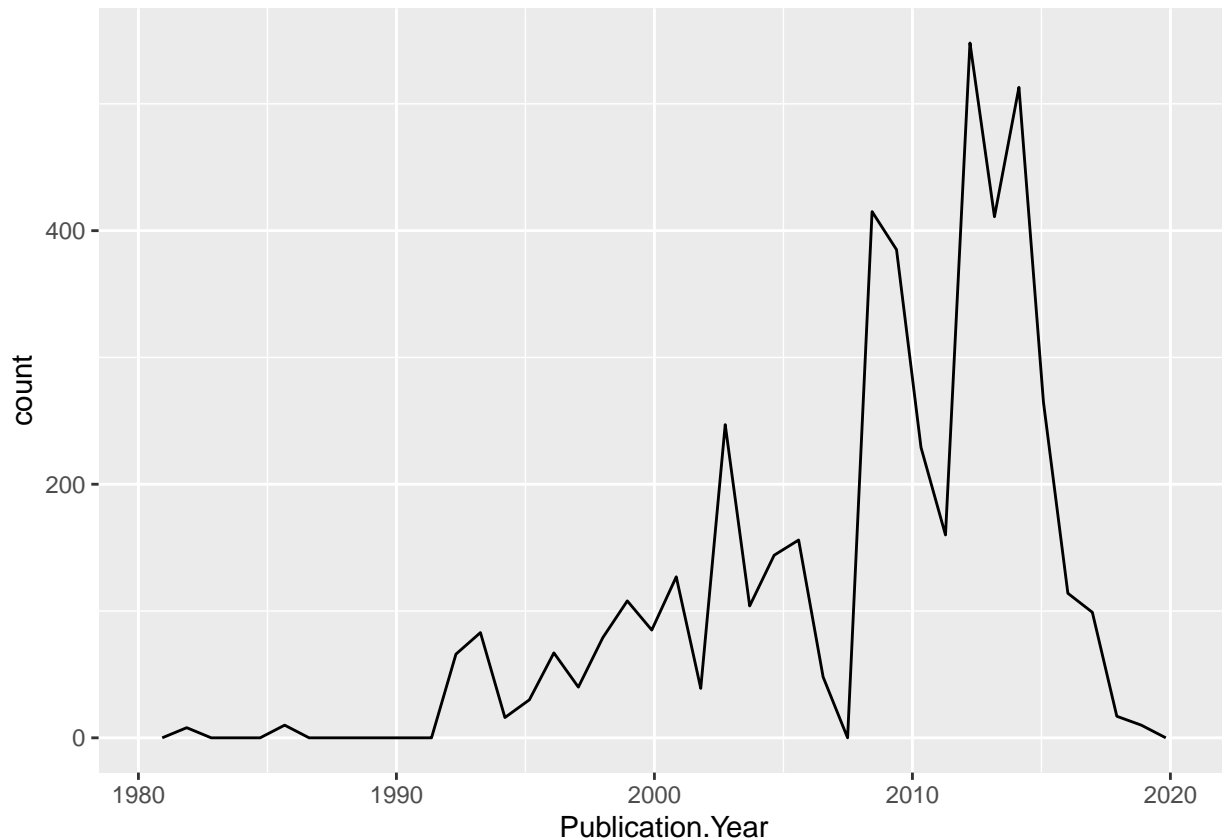
```
## [1] "factor"
```

Answer: The column above is a factor. It is likely not numeric because it is a form of nominal data - data used to categorize or classify entries (rather having quantitative significance). Also, since we told R to “import strings as factors,” if `Conc.1..Author` was a string in the CSV file it would’ve been converted to a factor upon running the command.

Explore your data graphically (Neonics)

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

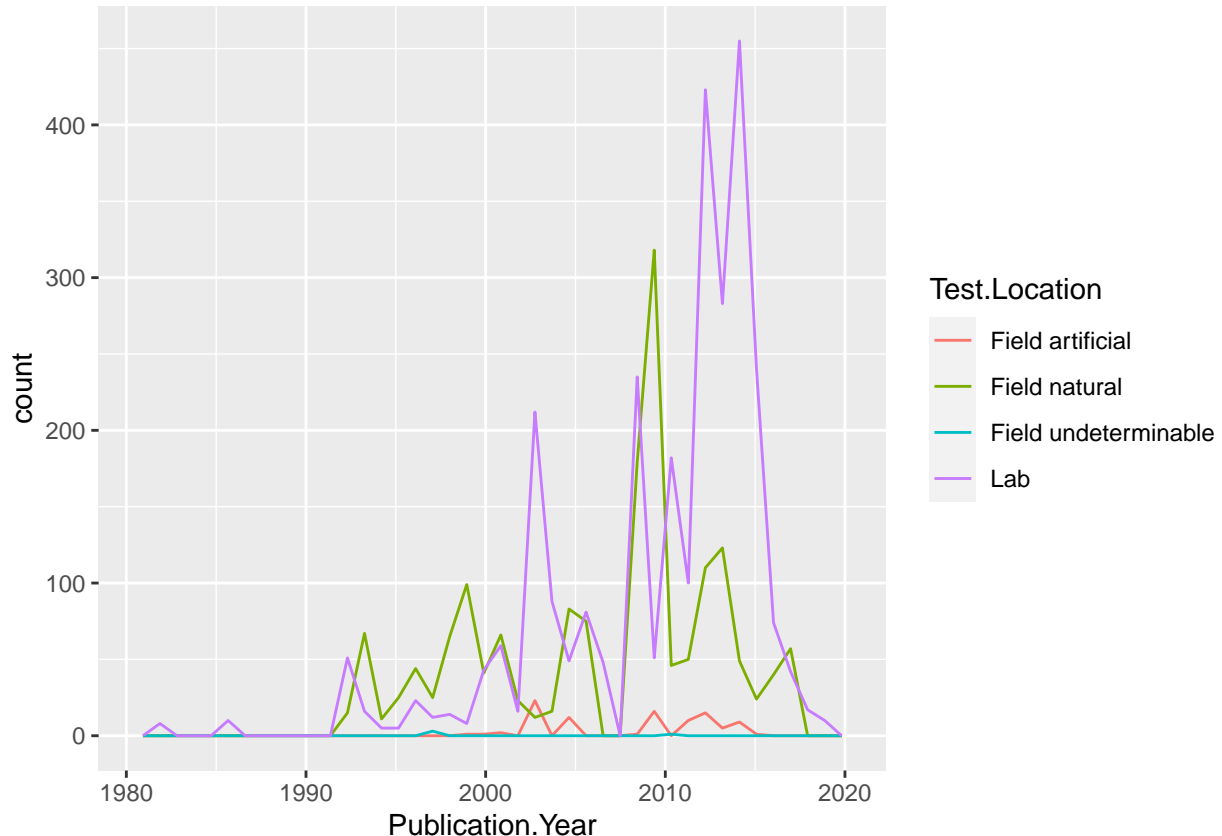
```
ggplot(Neonics) +  
  # using the ggplot package on the Neonics data set  
  geom_freqpoly(aes(x = Publication.Year), bins=40)
```



```
# plotting Publication.Year with bin width of 40
```

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

```
ggplot(Neonics) +  
# using the ggplot package on the Neonics data set  
geom_freqpoly(aes(x = Publication.Year,color=Test.Location), bins=40)
```



```
# plotting Publication.Year with bin width of 40; adding Test.Location and  
## displaying in different colors
```

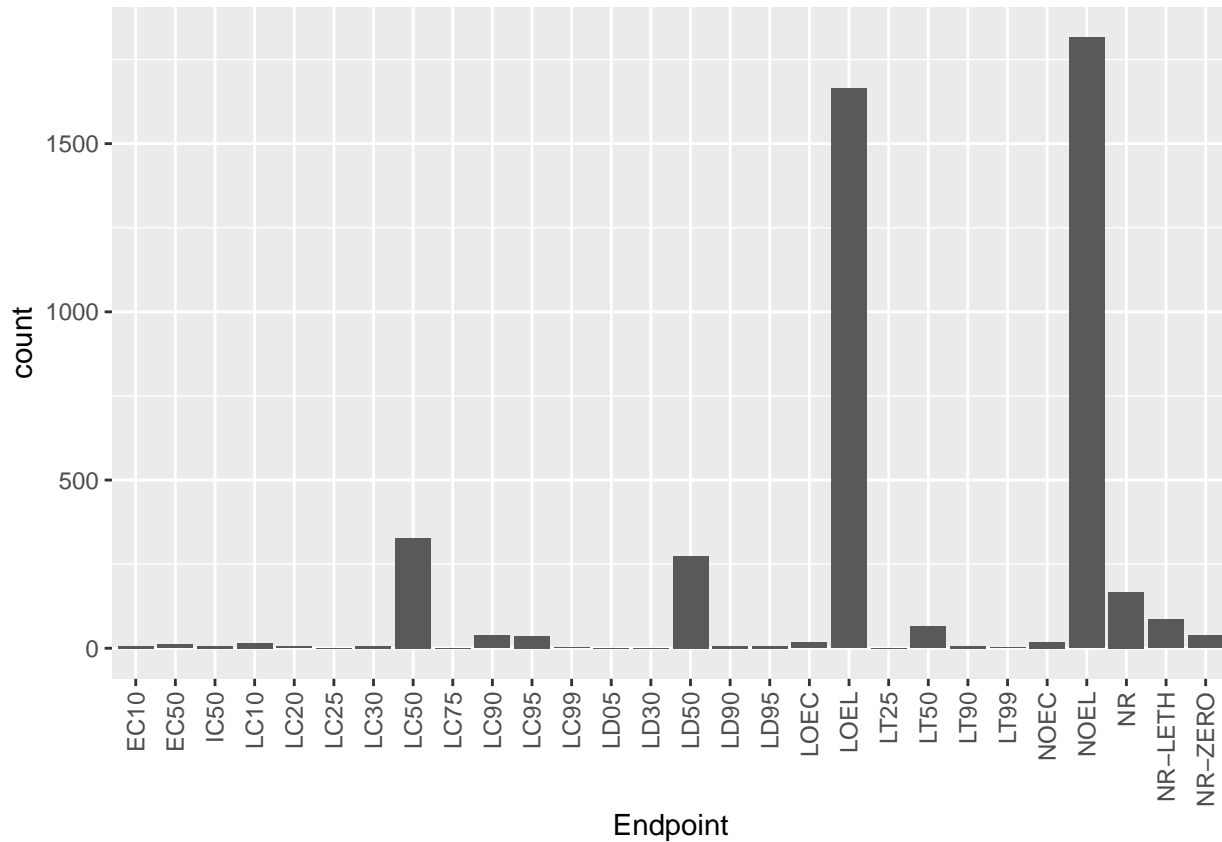
Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: The most common test locations are “Lab” and “Field natural.” They differ a bit over time, but “Lab” has been the most common for approximately 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, aes(x = Endpoint)) + # using ggplot to specify data set and column
  geom_bar() + # specifying type of graph as bargraph
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) # rotating
```



```
## x-axis labels for easier viewing
```

Answer: The most common end points are NOEL and LOEL. Per the appendix, NOEL is defined as “No-observable-effect-level” and represents :the highest dose (concentration) producing effects not significantly different from responses of controls according to author’s reported statistical test”. LOEL is defined as the “lowest-observable-effect-level” and represents the “lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls.”

Explore your data (Litter)

- 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) # class of collectDate is a factor
```

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



```
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
# converting collectDate column from factor to date using as.Date() function

class(Litter$collectDate) # column is now classified as date
```

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

```
unique(Litter$collectDate) # identifying unique dates sampled in August 2018
```

```
## [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`?

```
unique(Litter$plotID) # unique values in the plotID column. There were 12 unique plots sampled
```

```
## [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
```

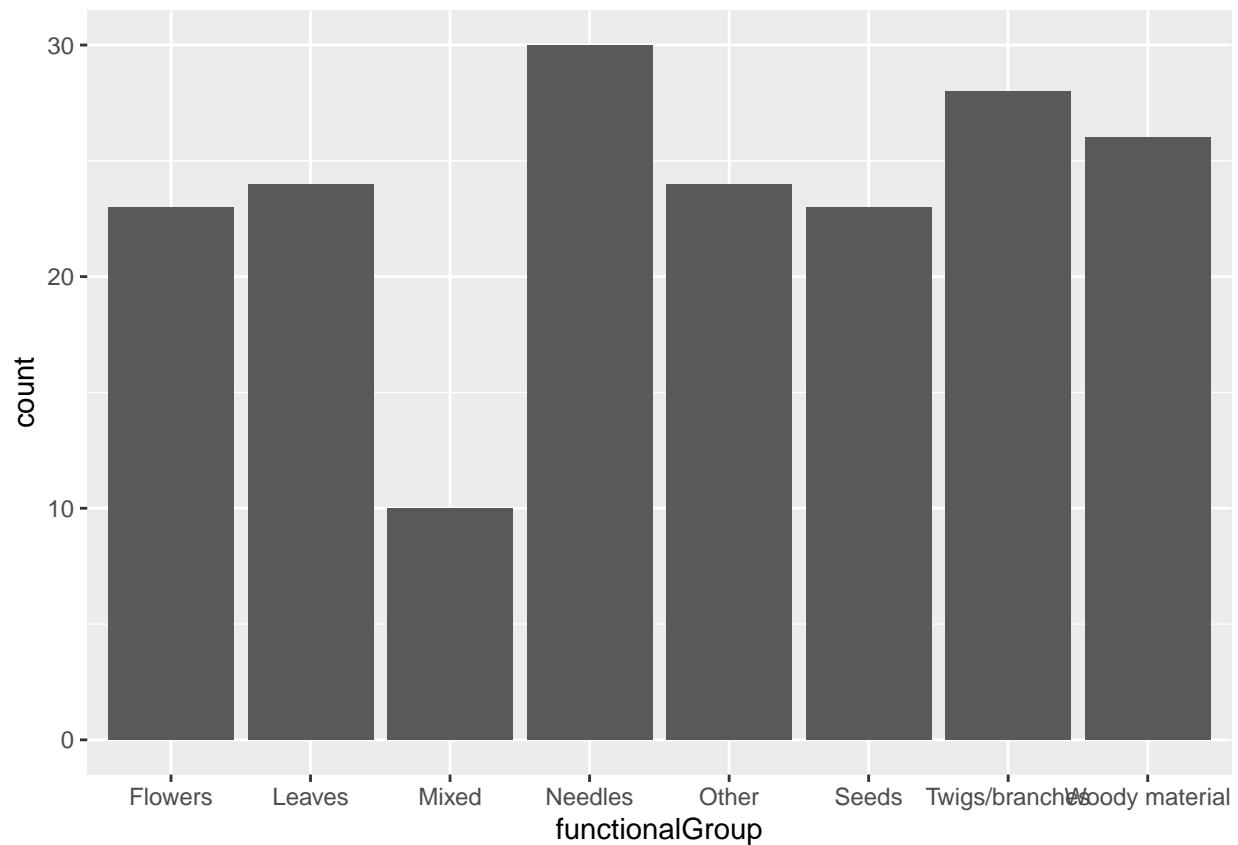
```
summary(Litter$plotID) # summary of values in the plotID column
```

```
## 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: Unique tells you the number of unique objects in the column while summary tells you how many of each unique object is in the column.

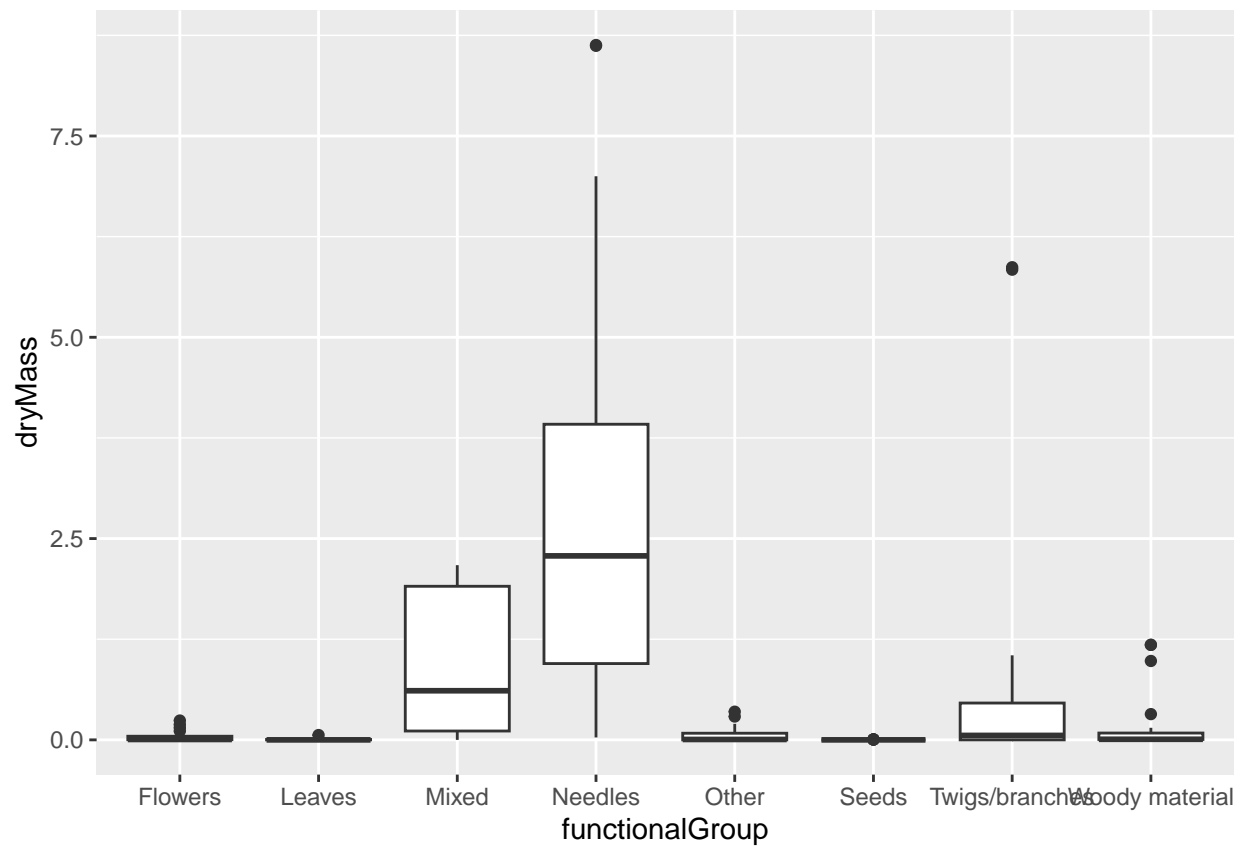
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(x = functionalGroup)) + # using ggplot to specify data set and column
  geom_bar() # specifying type of graph as bar graph
```

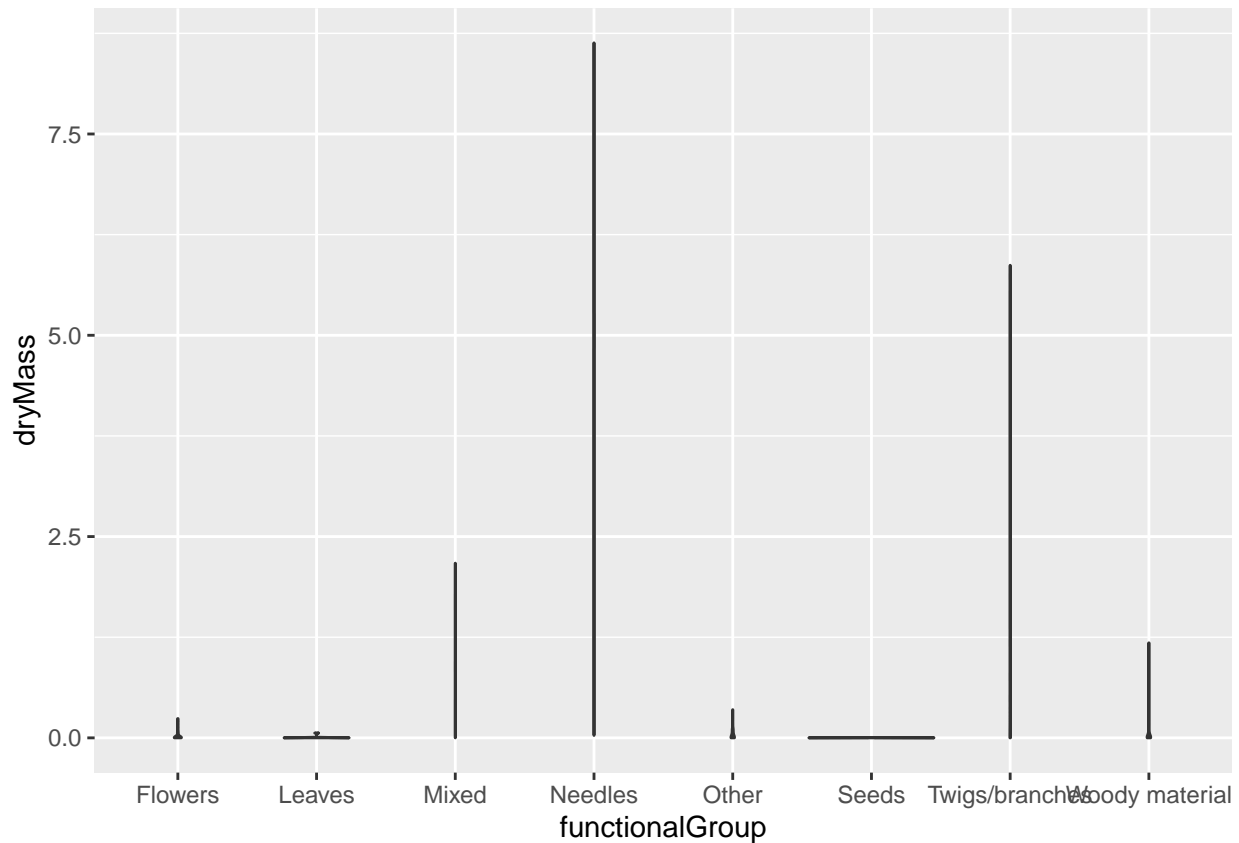


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of `dryMass` by `functionalGroup`.

```
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) + # inputs to ggplot
  geom_boxplot() # specifying type of graph as boxplot
```



```
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) + # inputs to ggplot
  geom_violin() # specifying type of graph as violin
```



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

Answer: The boxplot is a more effective option because the values in the data are very close together, so we are better able to see the distribution of data within a small range using the box plot. Since there is not a huge spread, it's difficult to discern anything meaningful about the data from the violin plot.

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

Answer: Needles tend to have the highest biomass as we can see from the boxplot.