

DwitiBagadia_A03_DataExploration

Dwiti Bagadia

Spring 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.

```
#load packages and set working directory
library(tidyverse)
library(lubridate)
getwd()
```

```
## [1] "/Users/d/Desktop/UNC/Spring:23/DUKE 872 L1 - Environmental Data Analysis/EDA-Spring23/Assignment"
```

```
setwd('/Users/d/Desktop/UNC/Spring:23/DUKE 872 L1 - Environmental Data Analysis/EDA-Spring23/Data/Raw')
```

```
#upload datasets
Neonics <- read.csv("ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = TRUE)
Litter <- read.csv("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: Since the 1990s, neonicotinoids have become one of the most commonly used insecticides in the U.S. and the world. They are now widely used in the agricultural sector as they possess a low mammalian toxicity and are effective for controlling stubborn pests in the soil. As neonicotinoids have proved to lead to loss of pollinators, they can give us a better understanding on the mechanisms behind the various diseases caused to the pollinators.

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 and the woody debris may act as a tinder that encourages the spreading of forest fires and in some cases also the start of the forest fire; in addition to their negative impacts they even act as a habitat for insects. Thus understanding about the quantity of the habitat for insects and influence on carbon budgets is important for integrating the ecotoxicity data.

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 fine woody debris are collected from elevated and ground traps. 2. The mass data is collected to an accuracy of 0.01 grams, for each group. 3. The randomness of trap placements is based on the type of vegetation.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#obtaining dimensions of NEONICS dataset  
dim(Neonics)
```

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

Answer: 4623(R). 30(C)

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?

```
#Summarizing the effects column  
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	

Answer: The most common effects of Neonics(Population and Mortality) help us understand about how common mortality is a result of neonicotinoids. It helps us by indicating the non-selective nature of the insecticide and the serious effect it has on the pollinators.

- 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...]

```
#creating a sorted summary of if the Species with common name
sort(summary(Neonics$Species.Common.Name), decreasing = TRUE, na.last = 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

Answer: The most commonly studied insects were - Honey Bee (667) and Parasitic Wasp (285). The honey bees being one of the most common insects which benefit from pollinators is the one that is studied the most. As Neonicotinoids are extremely dangerous to pollinators and benefitting insects, this data is important to study.

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 function to check the data type
class("Conc.1..Author")
```

```
## [1] "character"
```

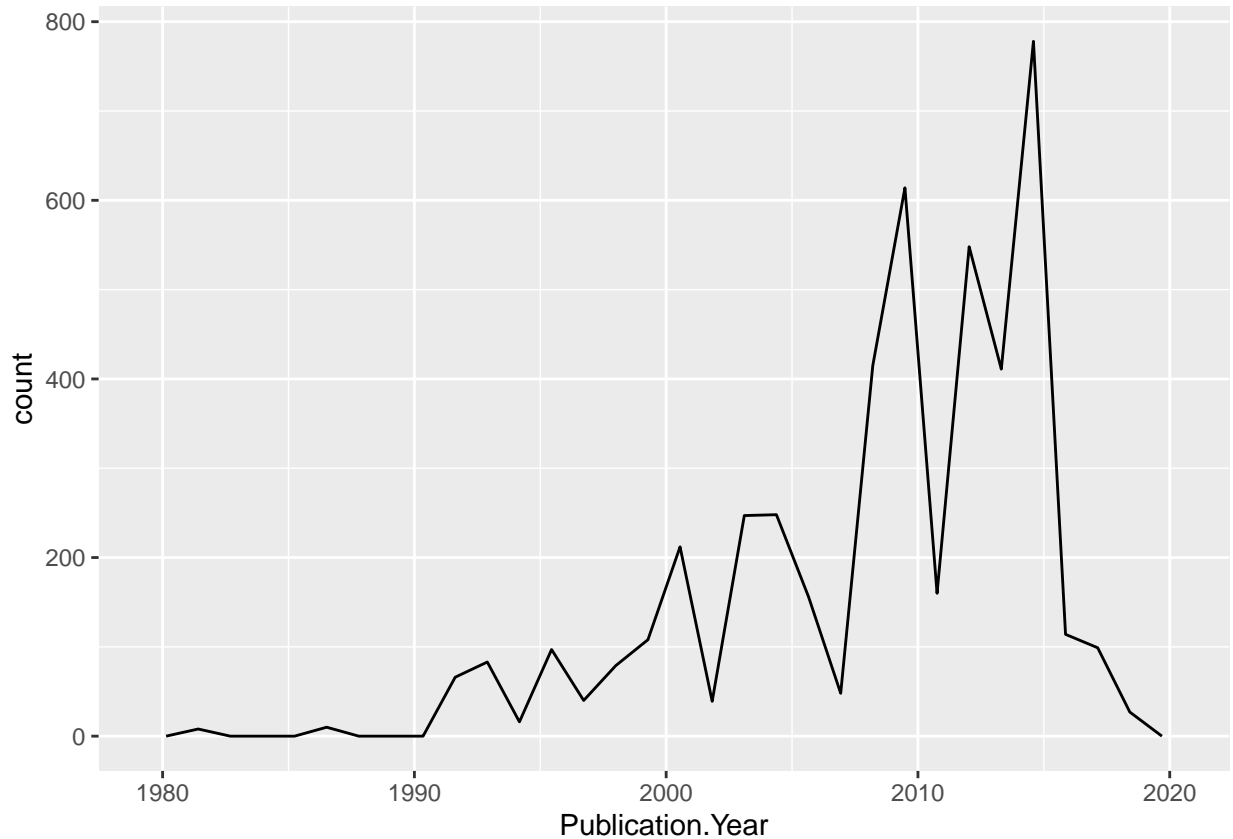
Answer: The `Conc.1..Author` is a character class, as the column has some data with both numbers and '<' signs. When some data in the column are characters, the class of the vector will convert to character instead of numeric.

Explore your data graphically (Neonics)

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

```
#creating a plot of number of studies done by publication year
ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year))
```

```
## '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.

```
#set of commands to plot the graph with colors
is.na(Neonics$Test.Location)
```

```
##      [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##     [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##     [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##     [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##     [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##     [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##     [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##     [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##     [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    [157] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    [169] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    [181] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    [193] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    [205] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    [217] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

[illegible]

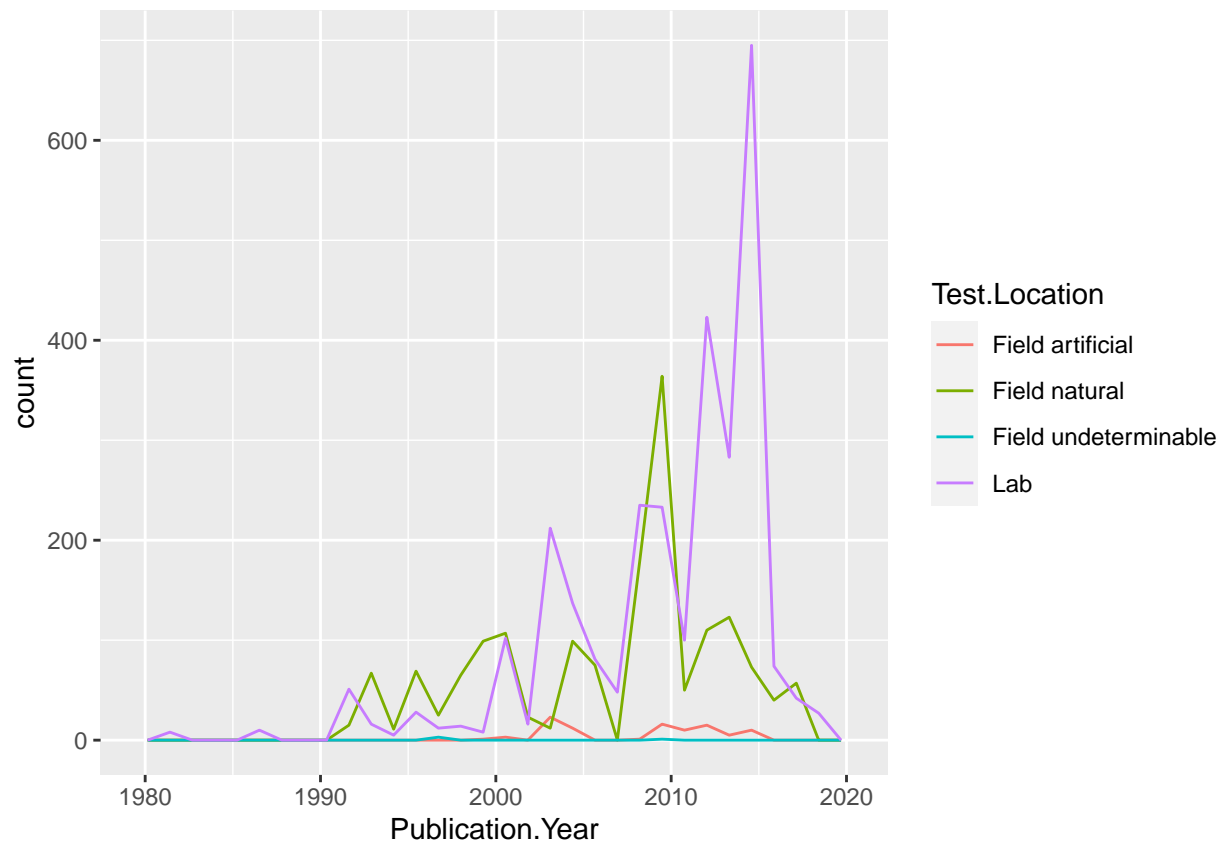
[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



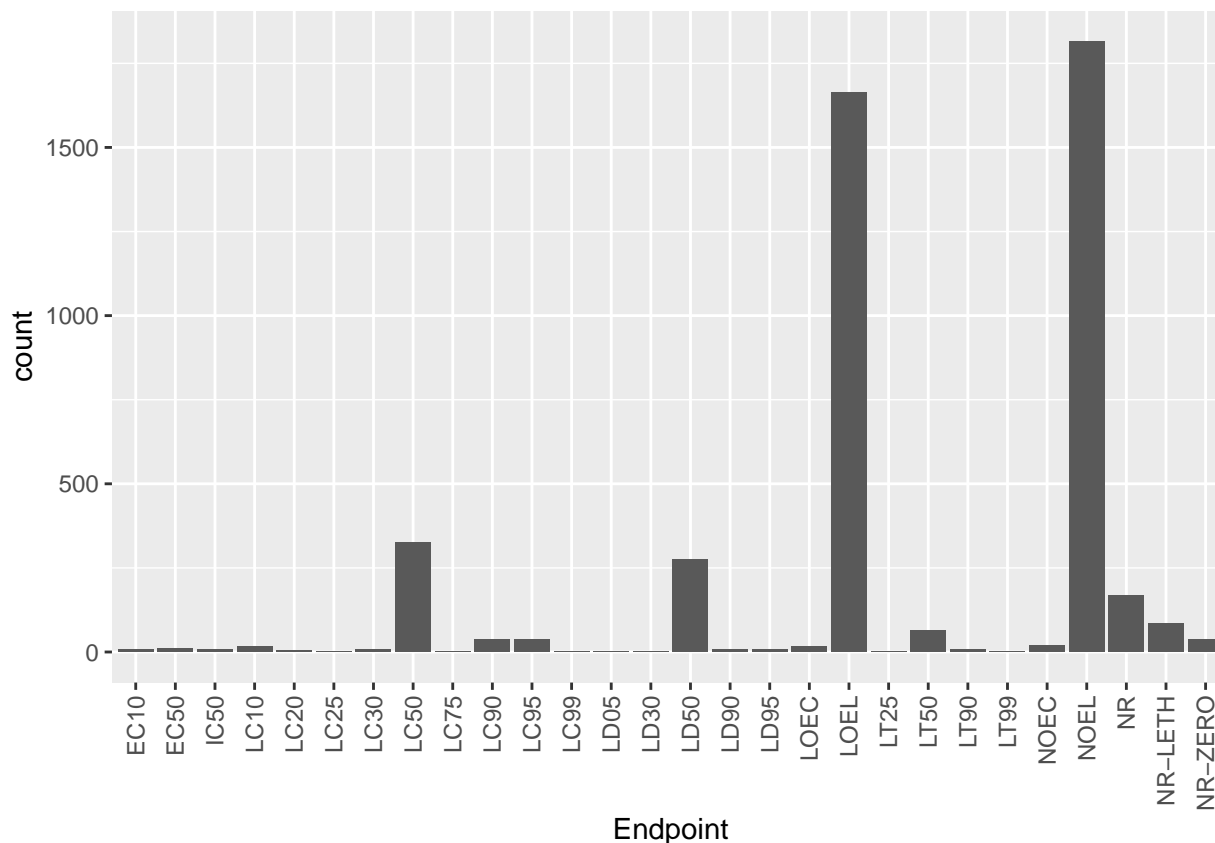
Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: Lab is the most test location more than once, first in 2000-2005 and then again in 2010-2020. While in time span of 1990-2000 and 2005-2010 the natural field was more common 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...]

```
#plotting endpoint counts
ggplot(Neonics) +
  geom_bar(aes(x = Endpoint)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



Answer: LOEL and NOEL are the most common end points. LOEL is the lowest observable effect level - having the lowest dose of insecticides producing effects that were significantly different from responses of controls. NOEL is the no observable effect level - having the highest dose producing effects that were not significantly different from the responses of controls according to the statistical tests reported by the author.

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.

```
#data class check
class(Litter$collectDate)
```

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

```
#output is factor#converting to date
Litter$collectDate <- as.Date(Litter$collectDate, format("%y/%m/%d"))
#checking class
class(Litter$collectDate)
```

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

```
#summary and check
summary(Litter$collectDate <- as.Date(Litter$collectDate, format = "%y%m%d"))
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NA's
##      NA      NA      NA    "NaN"     NA      NA    "188"
```

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

```
## [1] NA
```

```
#i cannot figure out what went wrong here.
```

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 for Niwot Ridge
unique(Litter$siteID)
```

```
## [1] NIWO
## Levels: NIWO
```

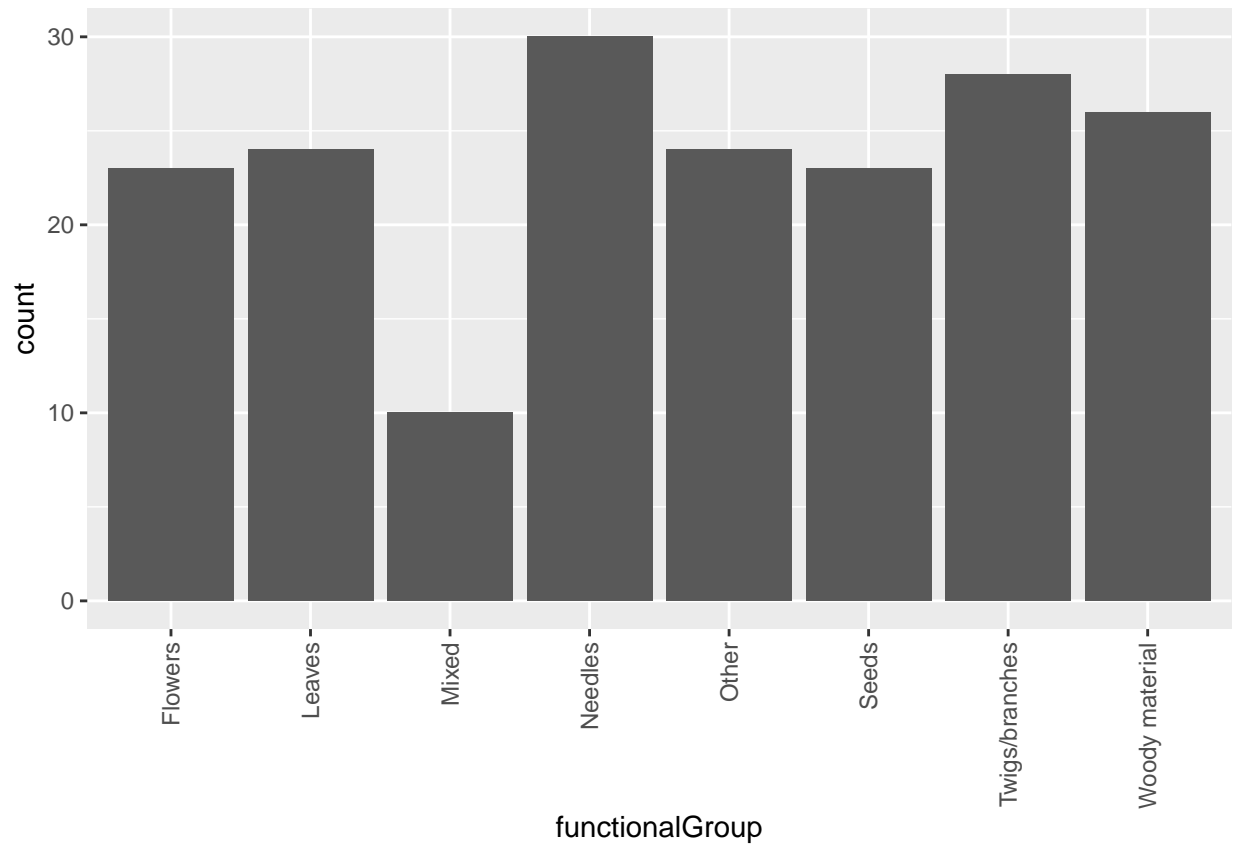
```
summary(Litter$siteID)
```

```
## NIWO
## 188
```

Answer: 188 plots were sampled at Niwot Ridge, `unique` command eliminated all the duplicate values, so the outcome was only 1 value - NIWO. where the `summary` command calculates the number of each value hence, giving us the number of the total plots (188) sampled at Niwot Ridge

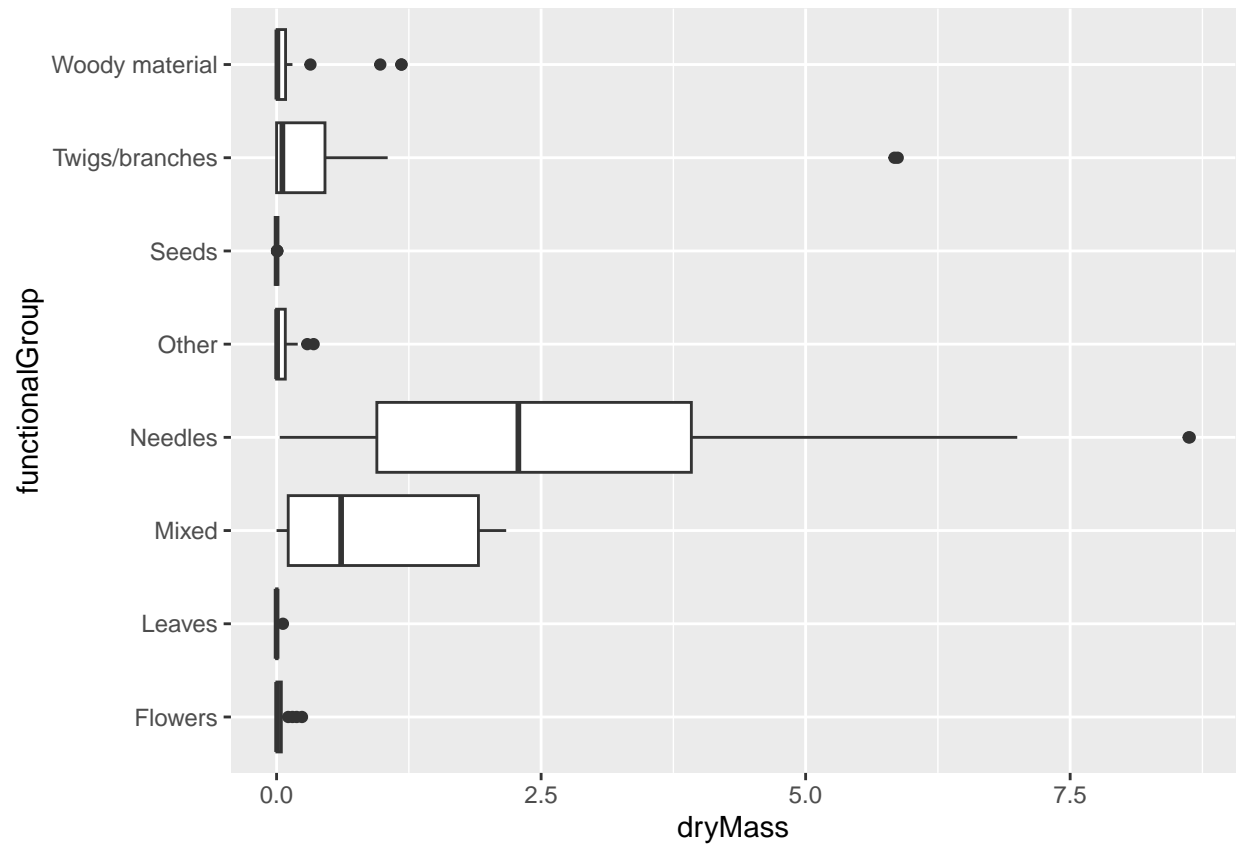
14. Create a bar graph of `functionalGroup` counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter) + geom_bar(aes(x=functionalGroup)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```

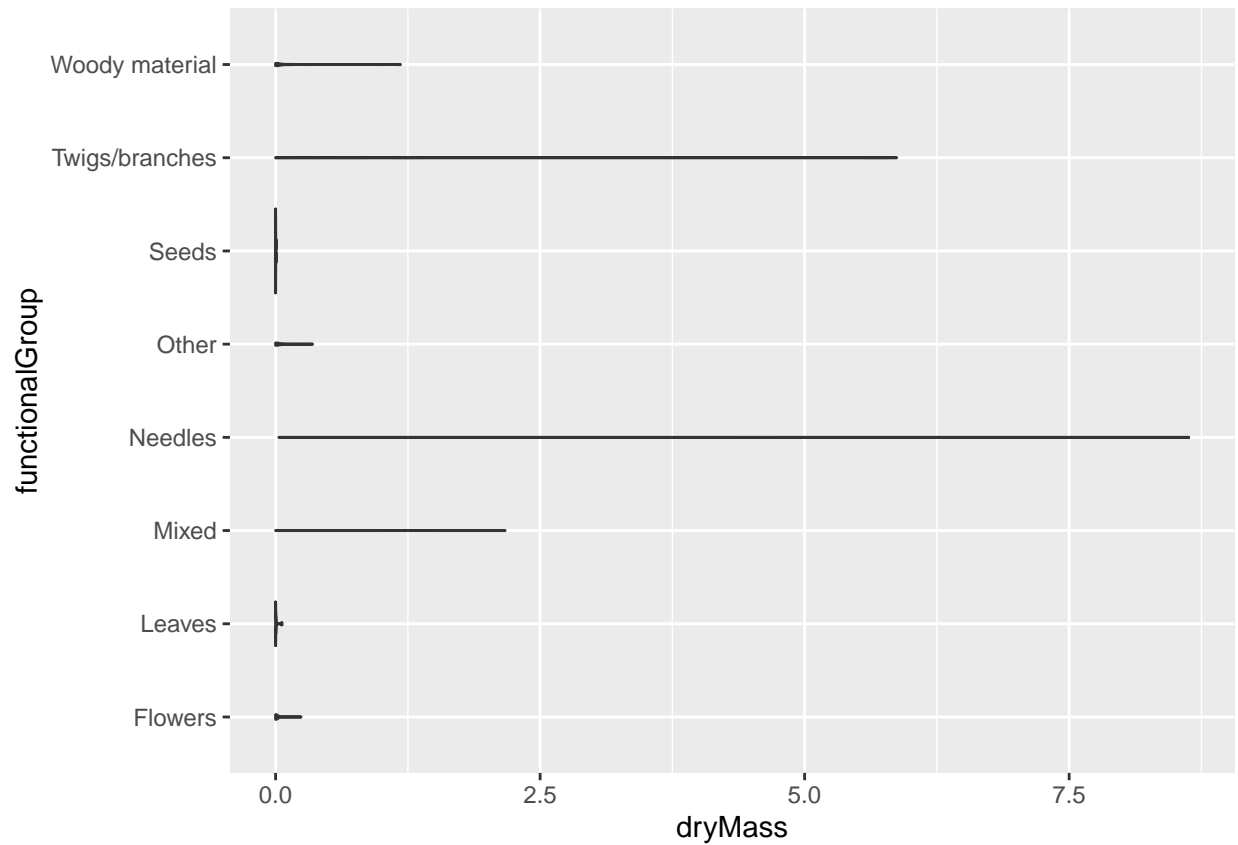



15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of `dryMass` by `functionalGroup`.

```
#box plot  
ggplot(Litter) + geom_boxplot(aes(x=functionalGroup, y=dryMass)) + coord_flip()
```



```
#violin plot  
ggplot(Litter) + geom_violin(aes(x=functionalGroup, y=dryMass)) + coord_flip()
```



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

Answer: In this case, each functional group more aesthetically and captures the viewer's attention to each group. The box plot shows only the summarized data, whereas the violin plot shows all the data.

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

Answer: Needles