

Assignment 3: Data Exploration

Queenie Wei

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/ziyaw/Downloads/EDE_Fall2023"
```

```
#loading necessary packages  
#install.packages("tidyverse")  
#install.packages("lubridate")  
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.1
```

```
## Warning: package 'lubridate' was built under R version 4.3.1
```

```
library(lubridate)
library(ggplot2)
#loading the two datasets and assigning them to Neonics and Litter, respectively
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",stringsAsFactors = TRUE)
Litter <- read.csv("./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: Because it would inform farmers if they should use this kind of insecticide, when to use it, how often they should use it, and for what kind of insects. In addition, farmers can learn about the negative impacts this may have on non-pests (pests are a made-up concept) such as bees and butterflies.

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: One of the most prominent uses for studying leaf litter is to study the decomposition rate at a certain site. This can in turn lead to many other fields of study – study of fungi, study of the differences in decomposition rates among plant species, the study of microclimates, etc. It is also a way scientists use to inform ecosystem function in rivers. The woody debris are also a great way to study carbon recycling in forest and aquatic ecosystems. It also constitutes habitats for different organisms, and helps with the structure and roughness of the habitat.

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. Trap placement may be targeted or randomized. 2. Ground traps are sampled once a year, and target sampling varies by vegetation type and cover. 3. Sampling only occurs in tower plots and are exclusively terrestrial.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#checking the dimensions of the two datasets
dim(Litter)
```

```
## [1] 188 19
```

```
#188rows 19columns
dim(Neonics)
```

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

```
#4623rows 30columns
```

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?

```
#showing how much
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
```

```
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
```

Answer: most common effects: Population, Mortality, Behavior, Feeding behavior, Reproduction. These effects are especially of interest because they are directly related to how the insects are affected by the neonicotinoids and the results of neonicotinoids (e.g. population and mortality).

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(Neonics$Species.Common.Name)
```

##	Honey Bee	Parasitic Wasp
##	667	285
##	Buff Tailed Bumblebee	Carniolan Honey Bee
##	183	152
##	Bumble Bee	Italian Honeybee
##	140	113
##	Japanese Beetle	Asian Lady Beetle
##	94	76
##	Euonymus Scale	Wireworm
##	75	69
##	European Dark Bee	Minute Pirate Bug
##	66	62
##	Asian Citrus Psyllid	Parastic Wasp
##	60	58
##	Colorado Potato Beetle	Parasitoid Wasp
##	57	51
##	Erythrina Gall Wasp	Beetle Order
##	49	47
##	Snout Beetle Family, Weevil	Sevenspotted Lady Beetle
##	47	46
##	True Bug Order	Buff-tailed Bumblebee
##	45	39
##	Aphid Family	Cabbage Looper
##	38	38
##	Sweetpotato Whitefly	Braconid Wasp
##	37	33
##	Cotton Aphid	Predatory Mite
##	33	33
##	Ladybird Beetle Family	Parasitoid
##	30	30
##	Scarab Beetle	Spring Tiphia
##	29	29
##	Thrip Order	Ground Beetle Family
##	29	27
##	Rove Beetle Family	Tobacco Aphid
##	27	27
##	Chalcid Wasp	Convergent Lady Beetle
##	25	25
##	Stingless Bee	Spider/Mite Class
##	25	24
##	Tobacco Flea Beetle	Citrus Leafminer
##	24	23
##	Ladybird Beetle	Mason Bee
##	23	22
##	Mosquito	Argentine Ant
##	22	21
##	Beetle	Flatheaded Appletree Borer
##	21	20
##	Horned Oak Gall Wasp	Leaf Beetle Family
##	20	20
##	Potato Leafhopper	Tooth-necked Fungus Beetle

##	20	20
##	Codling Moth	Black-spotted Lady Beetle
##	19	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
##	Araneoid Spider Order	Bee Order
##	17	17
##	Egg Parasitoid	Insect Class
##	17	17
##	Moth And Butterfly Order	Oystershell Scale Parasitoid
##	17	17
##	Hemlock Woolly Adelgid Lady Beetle	Hemlock Woolly Adelgid
##	16	16
##	Mite	Onion Thrip
##	16	16
##	Western Flower Thrips	Corn Earworm
##	15	14
##	Green Peach Aphid	House Fly
##	14	14
##	Ox Beetle	Red Scale Parasite
##	14	14
##	Spined Soldier Bug	Armoured Scale Family
##	14	13
##	Diamondback Moth	Eulophid Wasp
##	13	13
##	Monarch Butterfly	Predatory Bug
##	13	13
##	Yellow Fever Mosquito	Braconid Parasitoid
##	13	12
##	Common Thrip	Eastern Subterranean Termite
##	12	12
##	Jassid	Mite Order
##	12	12
##	Pea Aphid	Pond Wolf Spider
##	12	12
##	Spotless Ladybird Beetle	Glasshouse Potato Wasp
##	11	10
##	Lacewing	Southern House Mosquito
##	10	10
##	Two Spotted Lady Beetle	Ant Family
##	10	9
##	Apple Maggot	(Other)
##	9	670

```
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 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: Honey bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, italian Honeybee. These species are commonly studied because they are vital pollinators that are negatively affected by neonicotinoids. Because the insecticide can present itself in pollen and nectar, there is a direct threat to the pollinators, and they are toxic to them.

- 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.)
```

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

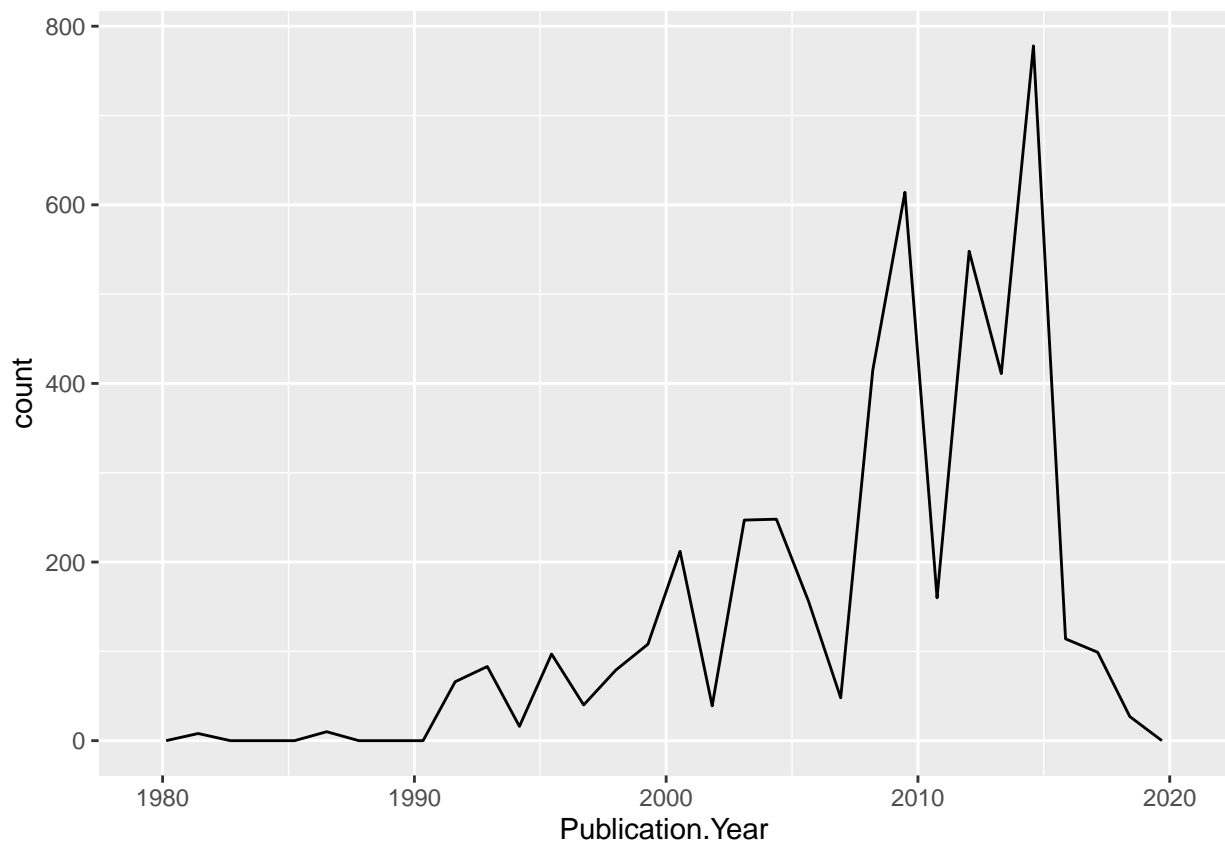
Answer: It is a factor because there are symbols (such as ~ and /) in some of the cells.

Explore your data graphically (Neonics)

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

```
#this generates a line plot for frequency in accordance with the  
#publication year  
ggplot(Neonics) +  
  geom_freqpoly(aes(x = Publication.Year))
```

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

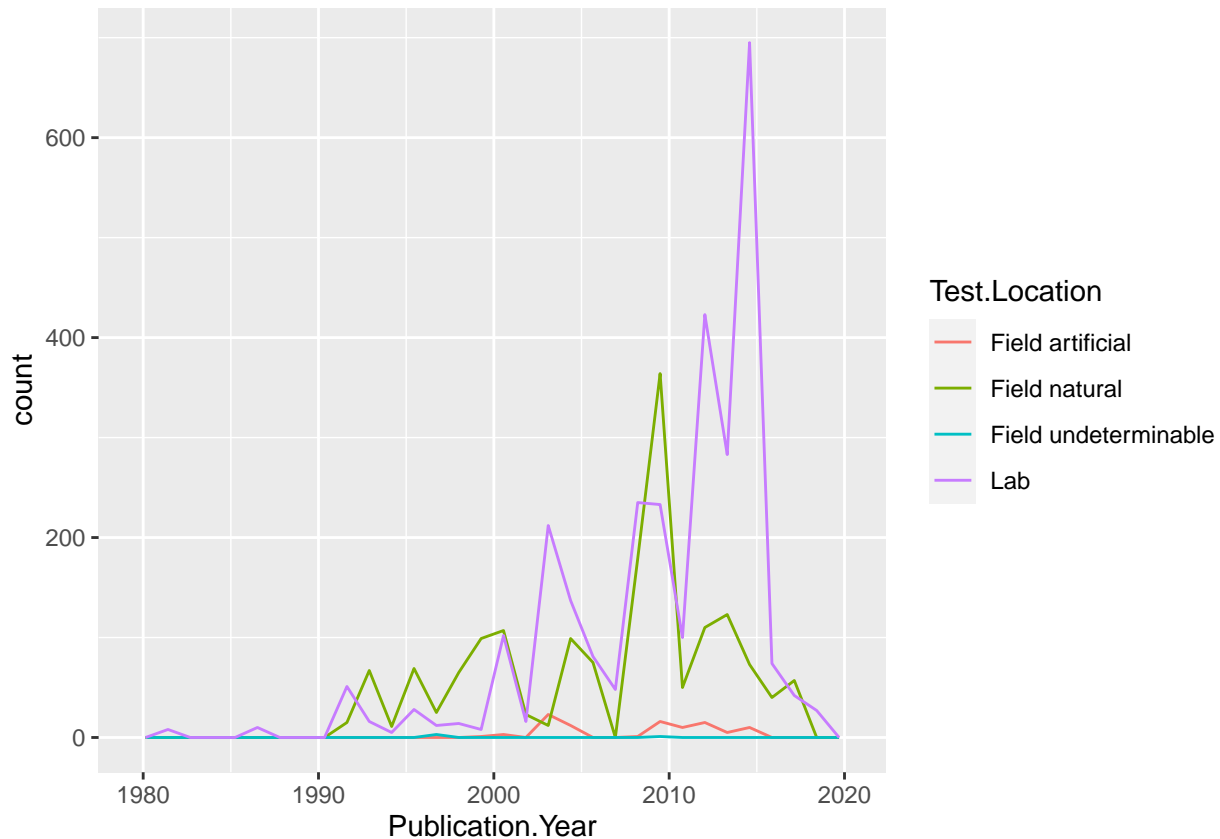


10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.


```
#this separates the line graph into several small ones according to their
#locations
```

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location))
```

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



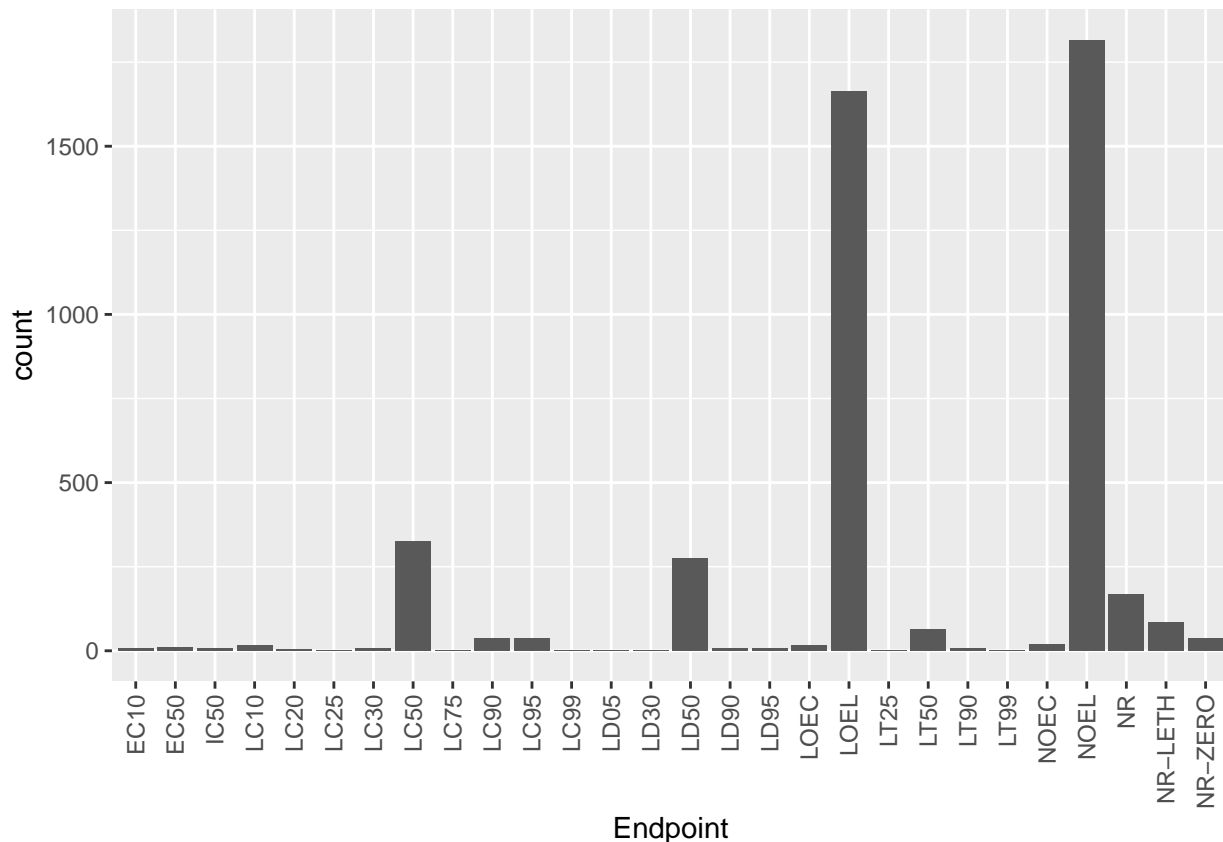
Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: the most common test locations would be in lab settings, although there was a peak of field natural locations around 2010. Tests are occurring a lot more in labs, and a lot less in field natural environments over time. In addition, there is a dropoff in all kinds of experiments around 2020, which may be attributed to the corona virus. There used to be more experiments in natural sites than lab sites, but with the advancement of technology, it is easier to control for different factors in labs, making it more favorable.

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))
```



Answer: The two most common end points are NOEL and LOEL, which stand for No-observable-effect-level and Lowest-observable-effect-level, respectively. LOEL is defined as the lower dose level that produces effects significantly different from responses to controls, and NOEL is the highest dose possible that does not produce a result that is significantly different from the 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)
```

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

```
#the class of collectdate is not date, it is factor.
```

```
today <- Sys.Date()
```

```
format(today, format = "%B")
```

```
## [1] "September"
```

```
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
class(Litter$collectDate)
```

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

```
#the class has been changed to Date
#finds the unique values in litter$collectdate
dates <- unique(Litter$collectDate)
dates
```

```
## [1] "2018-08-02" "2018-08-30"
```

```
#the dates were august 2nd and august 30th
```

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`?

```
length(unique(Litter$plotID))
```

```
## [1] 12
```

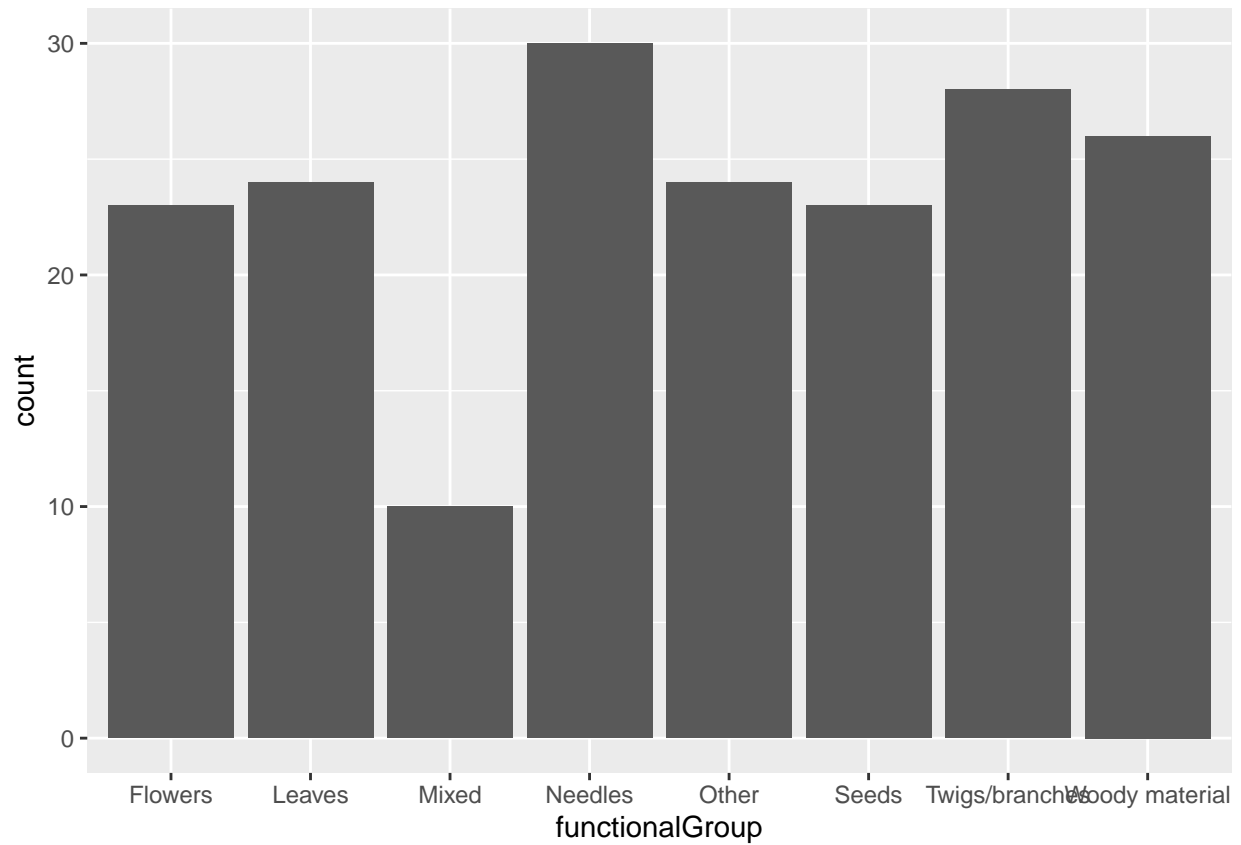
```
summary(Litter$siteID)
```

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

Answer: All 188 observations were collected at the Niwot Ridge. The information provided by the `unique` function does not automatically provide the number of unique values, whereas `summary` does. One would need to manually count the number of unique plots. `Summary` also provides other summary statistics when relevant.

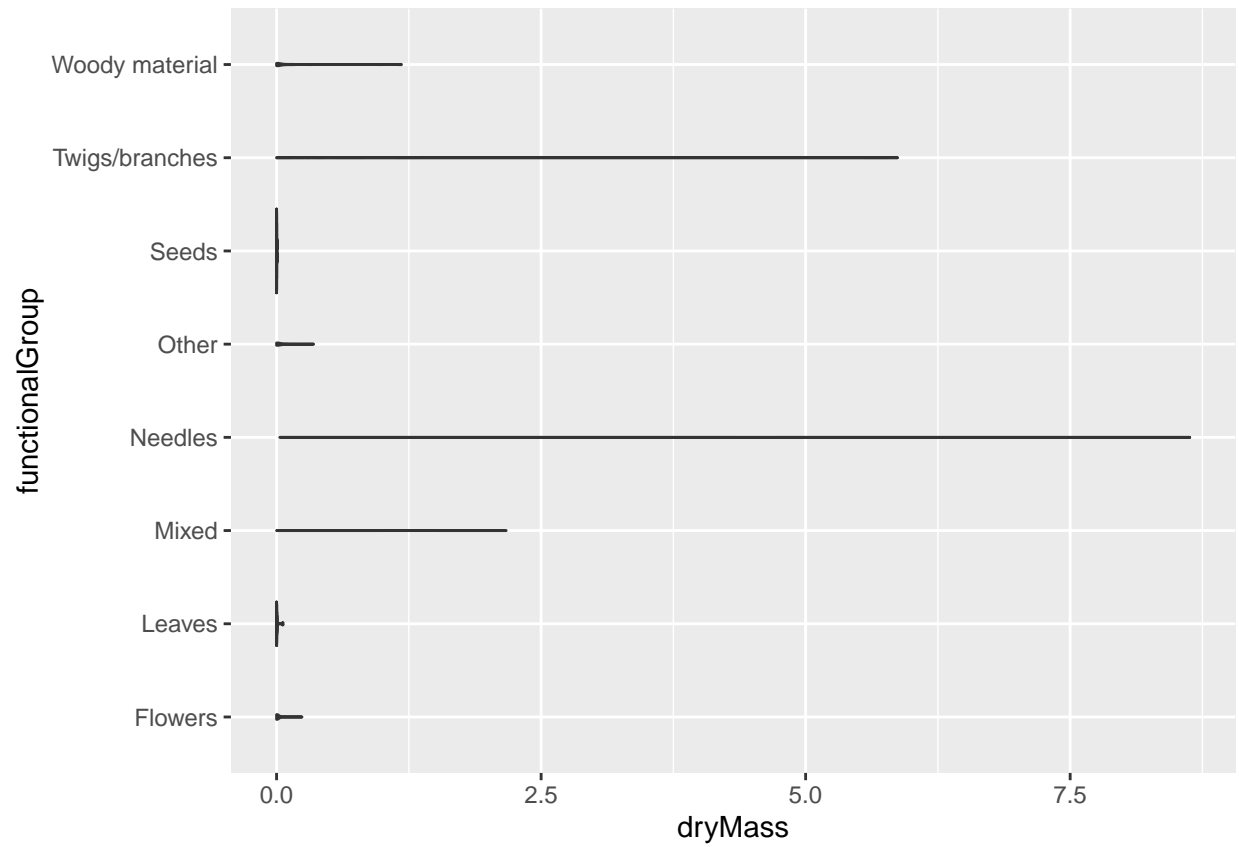
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter) +
  geom_bar(aes(x=functionalGroup))
```

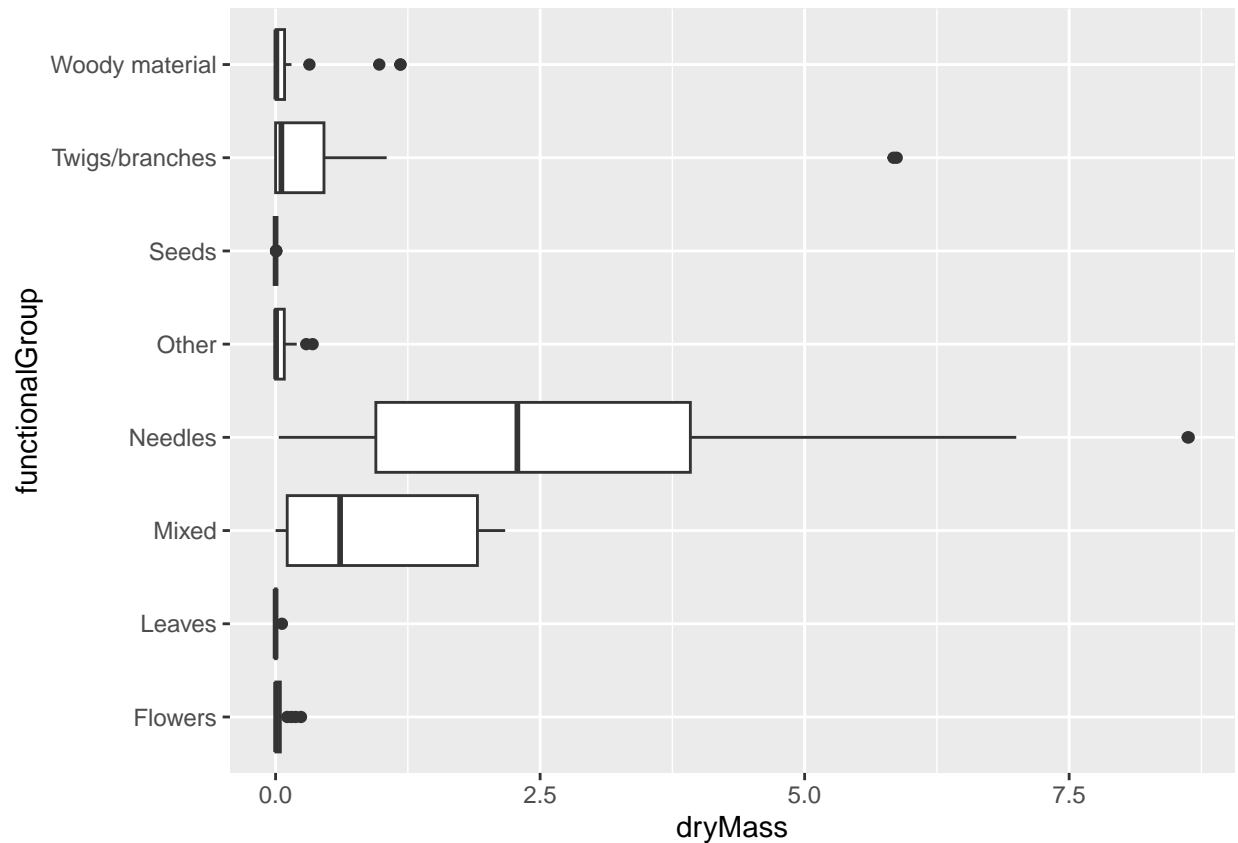


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of `dryMass` by `functionalGroup`.

```
ggplot(Litter) +  
  geom_violin(aes(x = dryMass, y = functionalGroup),  
    draw_quantiles = c(0.25, 0.5, 0.75))
```



```
ggplot(Litter) +  
  geom_boxplot(aes(x = dryMass, y = functionalGroup))
```



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

Answer: Because the data is very spread out with a lot of outliers. Boxplots are more robust when it comes to data with more outliers.

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

Answer: Needles tend to have the highest biomass at these sites according to the bargraph. We can derive this result by looking at the quartiles. The 2nd and 4th quartiles are much higher than the other functional groups.