SimonHeinberg_A03_DataExploration

Simon_Heinberg

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.

```
# 1. Set up working directory
getwd() #checking working directory
```

```
## [1] "/Users/simonheinberg/Desktop/Duke_Coursework/Envirnmental_Data_Exploration/EDE_Fall2023"
```

```
# 2. Load packages

library(lubridate)
library(tidyverse)
```

```r
# 3. Import datasets
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = T)
Litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv", stringsAsFactors = T)
```

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 neonicotinoid insecticides studies the impacts of insecticides on individual insects, population of insects, on organism communities, and on the broader ecosystem. This information is important to study in order to understand how the use of neonicotinoids might impact all of the above categories. Understanding the potential impacts of neonicotinoid-use will inform agricultural management and policy decisions. For example, certain neonicotinoids may be banned if they are found to have a sufficiently detrimental effect on the surrounding ecosystem.

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 woody debris on the forest floor have numerous impacts on a forest's ecosystem. Litter and woody debris play an important role in determining organic content in a forest's soil, which impacts soil health and can contribute to carbon sequestration. Litter and woody debris additionally impact a forest's susceptibility to wildfires. Litter and woody debris also help create habitat that supports life on the forest floor.

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 is collected by a pair of traps, namely one elevated trap and one ground trap. These pairs of traps are deployed for every 400 square meters of plot area. 2. Trap placement is either random or targeted depending on the percent of woody plant vegetation cover in the plot area. Trap placement is randomized for plots with greater than 50% covereage of qualifying woody vegetation. Trap placement is targeted for plots with less than 50% covereage of qualifying woody vegetation, with traps placed underneath the woody vegetation cover. 3. Ground traps are sampled once annually whereas elevated traps are sampled either once every two weeks or once every one to two months depending on forest type.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
NeoDim<-dim(Neonics)

NeoDim #reading the dimensions of "Neonics" dataset
```

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

```
#The "Neonics" dataset 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?

```
NeoEffect<-summary(Neonics$Effect)
NeoEffect #Summarizing the most common cells in the "Effect" column of the 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 common effects listed in order are 1. Population 2. Mortality and 3. Behavior. Understanding the most common effects of neonicotinoids on insects is important for performing any meta-analysis on the overall impacts of neonicotinoids. Neonicotinoids can have a variety of effects on insects, as the data lists 19 different effects. However, population and mortatlity effects are far more common than the other effects. This data helps inform an overall understanding of how neonicotinoids tend to affect insects, which is through population and mortatlity.

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...]

```
NeoSpecies<-summary(Neonics$Species.Common.Name)
NeoSpecies #Summarizing the six most commonly studied species
```

```
##                          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 Wooly 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
```

Answer: The six most common species, listed in order are 1. Honey Bee 2. Parasitic Wasp 3. Buff Tailed Bumblebee 4. Carniolan Honey Bee 5. Bumble Bee and 6. Italian Honeybee. Five of the six species are bees, which are pollinators. As pollinators, these species play an important role in ecosystem function and an important role in the reproduction of crops. The sixth species, parasitic wasp, may play an important role in controlling populations of agricultural pests. All six of these species likely provide benefits to agriculture production.

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?

```
Author_Class<-class(Neonics$Conc.1..Author.)
Author_Class #Determining the class of 'Conc.1..Author'
```

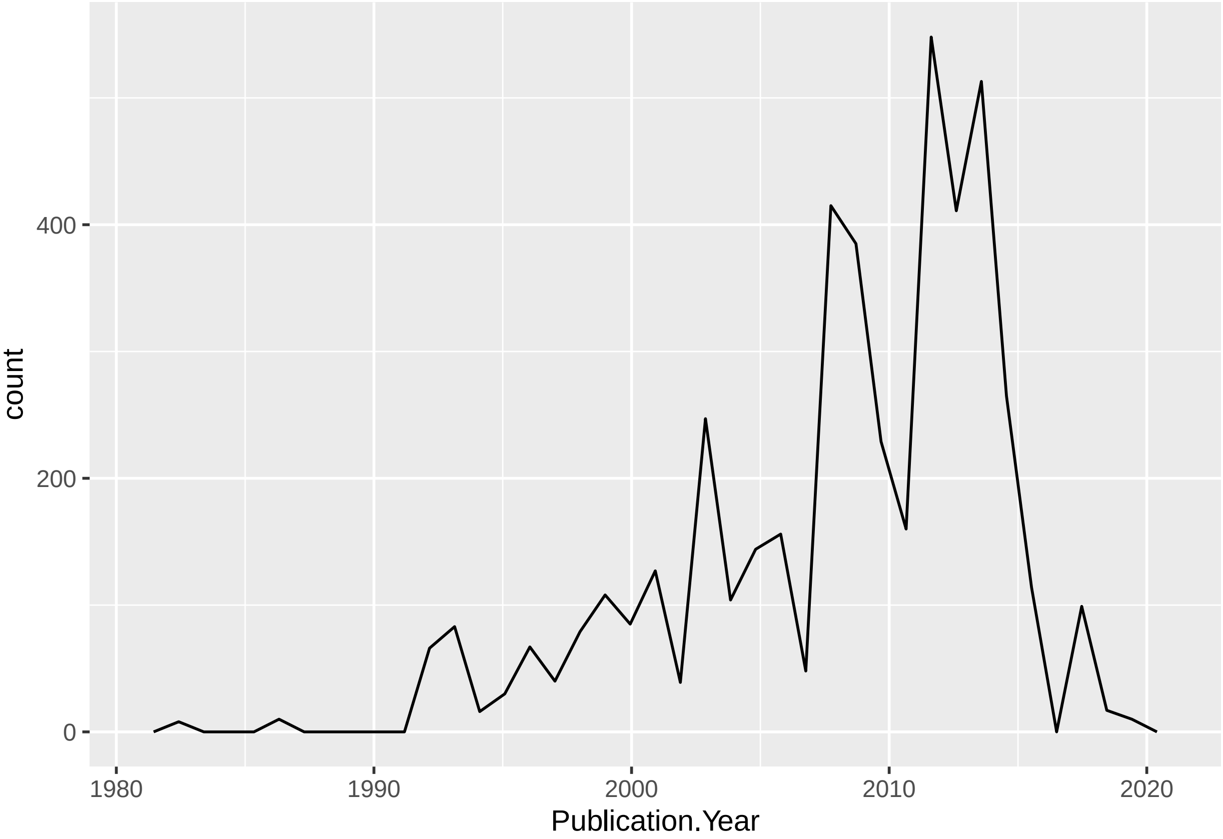
```
## [1] "factor"
```

Answer:The concentrations are a factor class. When importing the dataset, I directed R to read strings as factors. Since the 'Conc.1..Author' column has some cells with non-numeric values, such as 'NR/', it must have read the column as a factor.

Explore your data graphically (Neonics)

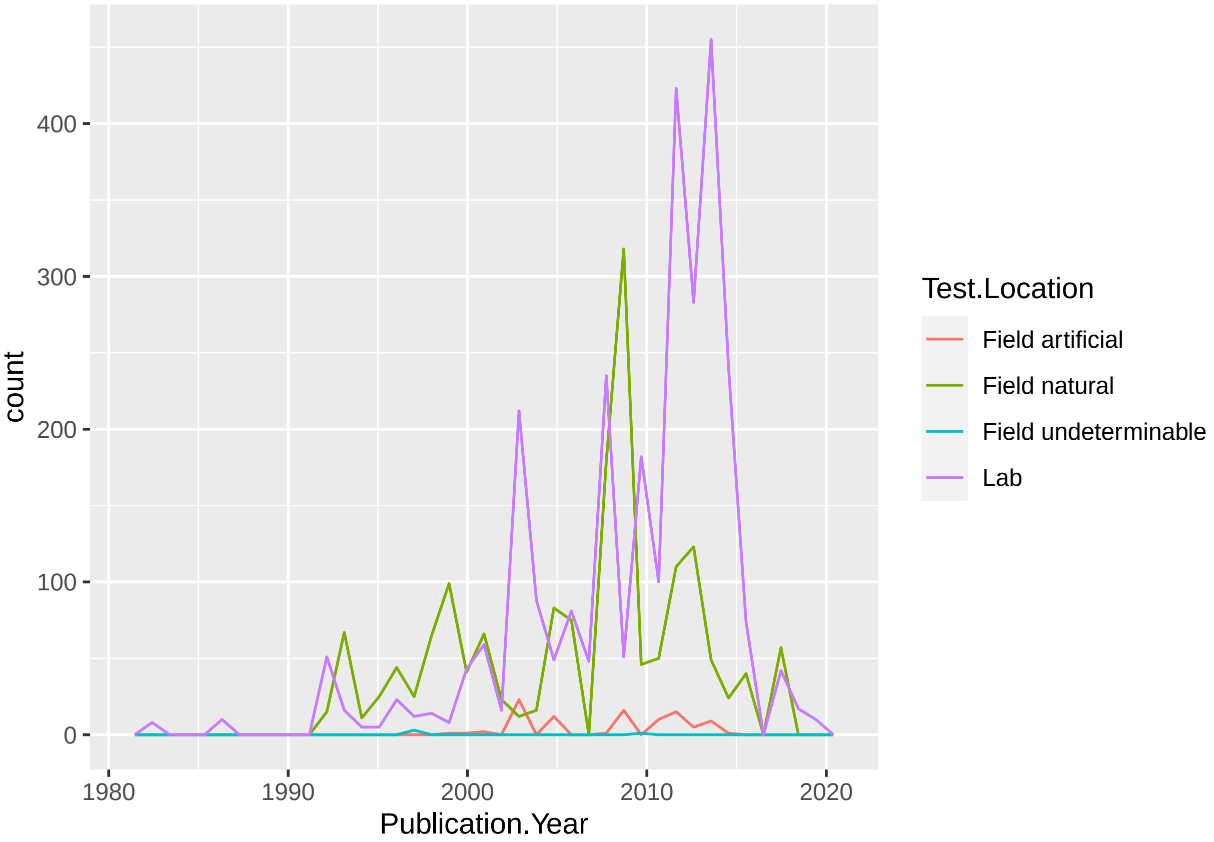
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
Studies_By_Year <-ggplot(Neonics)+geom_freqpoly(aes(x=Publication.Year), bins=39)

Studies_By_Year #generating a plot of the number of studies by publication year, which range from 1982-
```



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

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

Studies_By_Year_Colored #reproducing the graph with a different colored line for each 'Test.Location' v
```



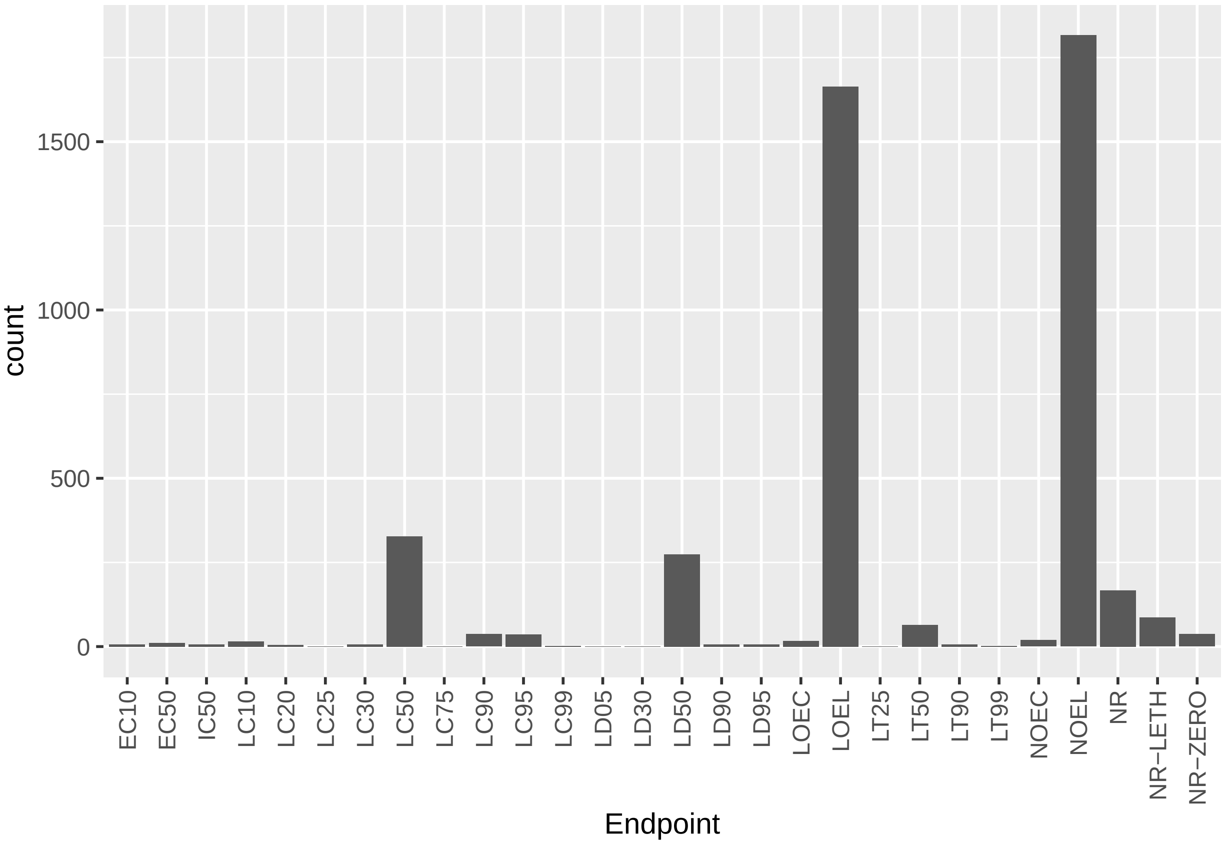
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer:The most common test locations are ‘Lab’. The ‘Lab’ test location increased in frequency after 2000, and in particular increased in frequency after 2010. In fact, all of the test locations increased in count after 2000 and in particular after 2010.

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…]

```
Endpoint_counts<-ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar()+theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

Endpoint_counts #creating a bar graph of Endpoint counts
```



Answer: The two most common Endpoint counts are NOEL and LOEL. The NOEL endpoint stands for 'no-observable-effect-level', which means that the highest concentration of insecticide did not produce a different effect than the control value. The LOEL endpoint stands for 'lowest-observable-effect-level', which means that the lowest dose of insecticide produced effects that were significantly different than 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.

```
collectDate_class<-class(Litter$collectDate)

collectDate_class #determining the class of 'collectDate'
```

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

```
#Result: collectDate is a factor and not a date

Litter$collectDate<-ymd(Litter$collectDate) #changing the 'collectDate' class from factor to date

August_Sampling<-unique(Litter$collectDate)

August_Sampling #determining which dates sampling occured during the month of August, 2018
```

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

```
#Sampling occured on two dates in August of 2018: 2018-08-02 and 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`?

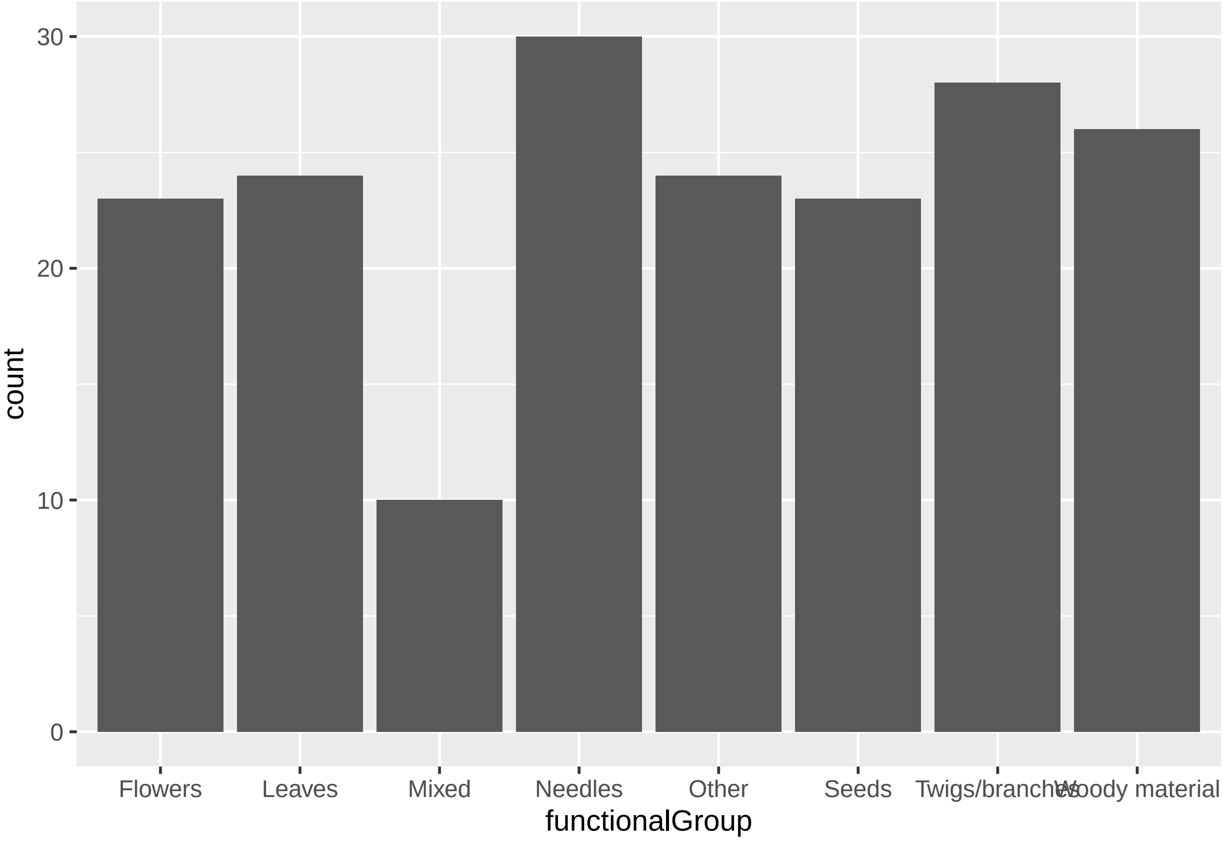
```
Unique_Plots<-unique(Litter$namedLocation)

Unique_Plots #will output the count of unique plots
```

```
##  [1] NIWO_061.basePlot.ltr NIWO_064.basePlot.ltr NIWO_067.basePlot.ltr
##  [4] NIWO_040.basePlot.ltr NIWO_041.basePlot.ltr NIWO_063.basePlot.ltr
##  [7] NIWO_047.basePlot.ltr NIWO_051.basePlot.ltr NIWO_058.basePlot.ltr
## [10] NIWO_046.basePlot.ltr NIWO_062.basePlot.ltr NIWO_057.basePlot.ltr
## 12 Levels: NIWO_040.basePlot.ltr ... NIWO_067.basePlot.ltr
```

> Answer:There are 12 unique plots. The 'unique' function outputs the number of unique plots whereas the 'summary' function outputs the number of occurences of each plot. Using the 'summary' function to identify the number of unique plots would require that one manually count the number of unique plots. This is why the 'unique' function works better for identifying how many unique plots there are.

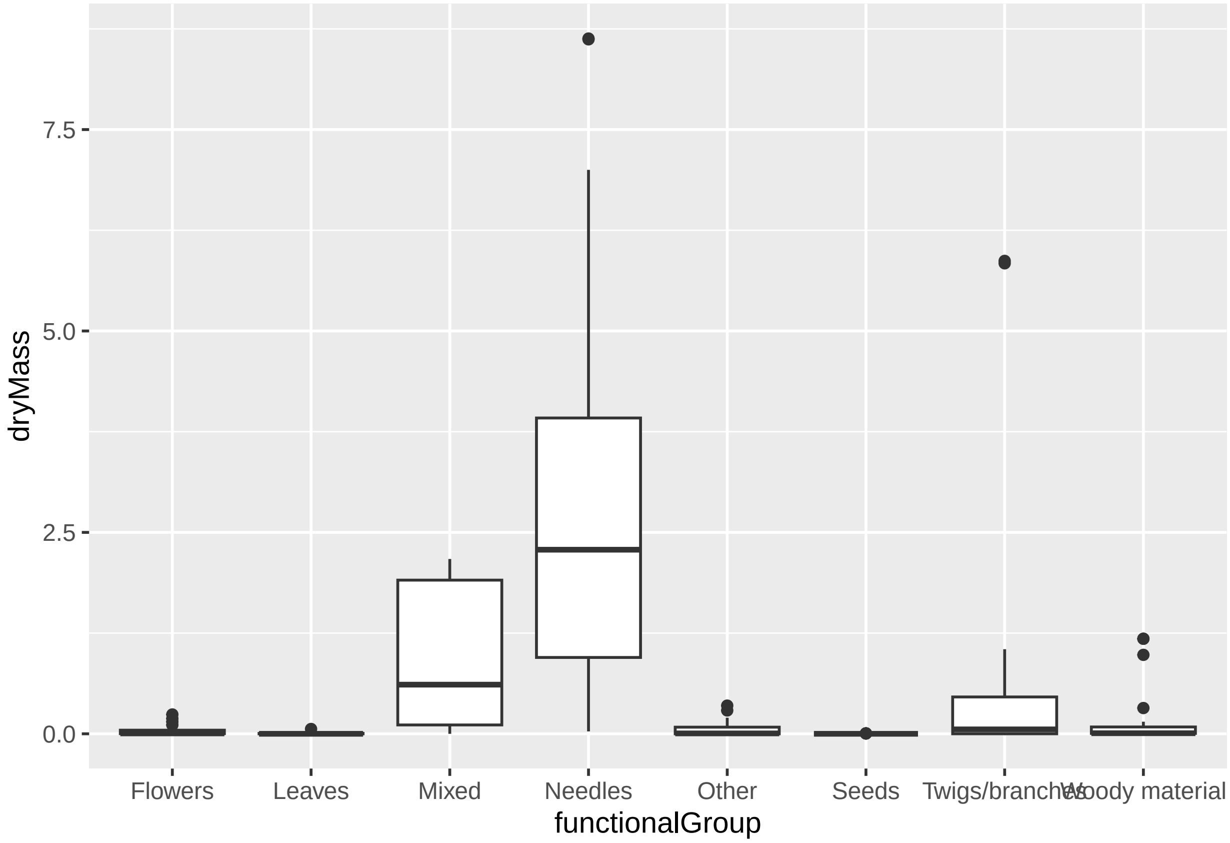
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
Litter_Types<-ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar()

Litter_Types #creating a bar graph of functionalGroup counts
```

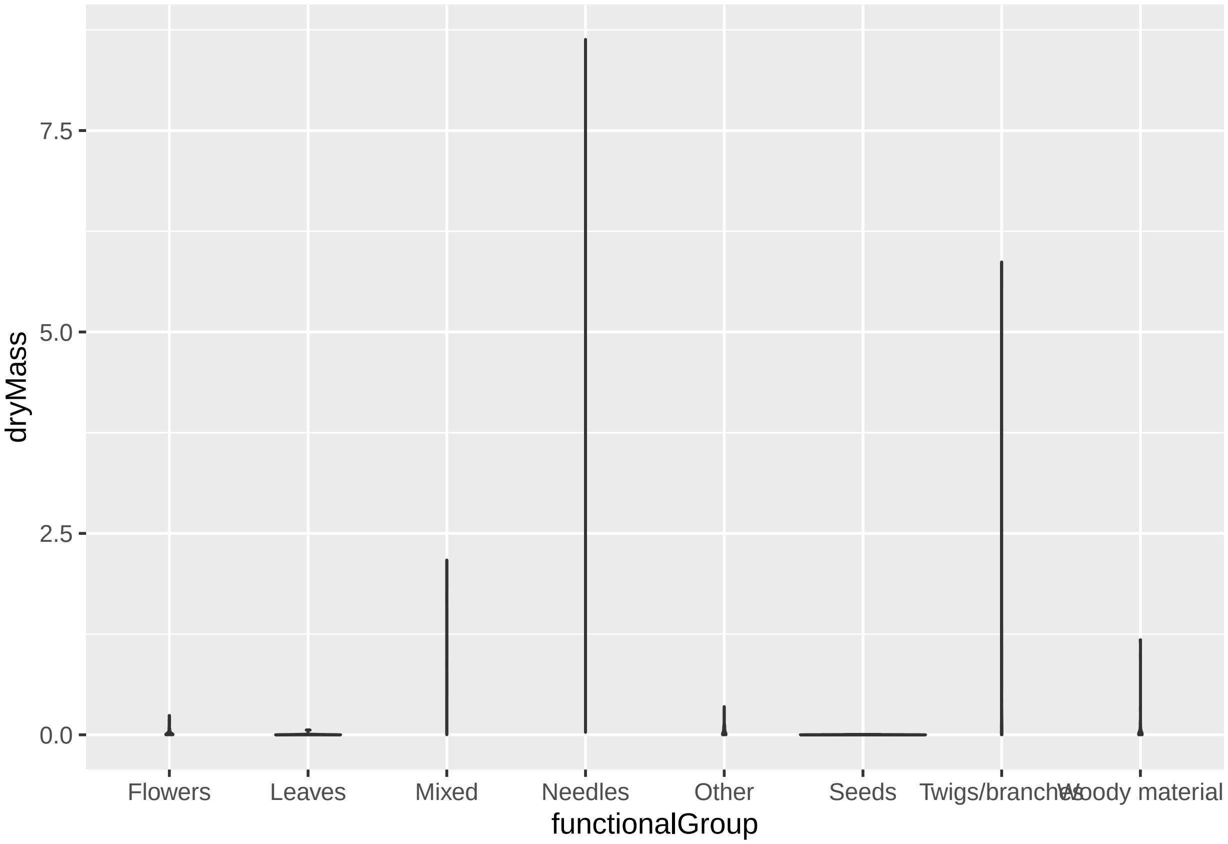


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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

Dry_Mass_Box #creating a boxplot of dryMass by functionalGroup
```



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

Dry_Mass_Violin #creating a violin plot of dryMass by functionalGroup
```



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

> Answer:The scale of the y-axis of the graph is skewed by the 'needles' category. As a result, it is difficult to see the distribution of several categories that have lower biomass values, namely the 'flowers', 'leaves', 'other', and 'seeds' categories. Of the categories that are easier to visualize, the distribution of the mass values is not concentrated around a given value. They appear to be nearly evenly distributed across a wide range of dryMass values. This means that on the violin plot, their graphs appear as vertical lines without visible median or IQR values. However, the box plot manages to show the IQR range and median for several of the litter types. These include the 'needles', 'mixed', and 'twigs/branch' categories.

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

> Answer:The 'needles' tend to have the highest biomass, with the highest median and IQR values. The 'mixed' litter has the second highest biomass as measured by the median.