Assignment 3: Data Exploration

Emily Kuhlmann

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.

```
getwd()
```

```
## [1] "/Users/emilykuhlmann/Documents/Documents/Fall23/EDA/EDE_Fall2023"
```

```
#loading the necessary packages
library(tidyverse)
library(lubridate)
library(ggplot2)
library(here)
```

```
#uploading the datasets with here()
Neonics <- read.csv(file = here('Data','Raw','ECOTOX_Neonicotinoids_Insects_raw.csv'),
                    stringsAsFactors = T)
Litter <- read.csv(file = here('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: Understanding the ecotoxicology of neonics provides information on application rates and amounts, because the different toxicological endpoints indicate how much of the compound is needed to cause the desired outcome in each organism. Additionally, the toxicology can provide insight into what nontarget species may be impacted by application of the insecticide. This is important in reducing the ecological impacts of insecticide application.

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: Studying litter and wood debris can provide important information about the nutrients that will be entering the soil as the debris decays. Litter and woody debris also provide habitats for organisms, so understanding its composition can provide information about this ecosystem.

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 from elevated traps and woody debris are collected from ground traps. 2. Mass data are measured for separate functional groups which include leaves, needles and seeds. 3. There are many specifications in the spatial sampling setup, including that plot edges must be separated by a distance that is 150% of one edge of the plot.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
Neonic_dim <- dim(Neonics) #dataset dimensions
Neonic_dim
```

```
## [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?

```
Common_effect <- summary(Neonics$Effect)
#summarizing the effects, lists the count for each effect
Common_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 are population and mortality. These are of specific interest because neonicotinoids are used to eradicate pests in agriculture, so knowing how the compounds impact the population size and mortality of organisms will demonstrate how effective the compound is for that organism.

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. . .]

```
Common_Name <- sort(summary(Neonics$Species.Common.Name), decreasing = T)
#summarizing the species and sorting in order of decreasing frequency
Common_Name
```

```
##                          (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 six most commonly studies species (besides 'other') are Honey Bee, Parasitc Wasps, Buff Tailed Bumblebees, Carniolan Honey Bees, Bumble Bees, and Italian Honeybees. These are widely studied because they are pollinators, so it is important to try and reduce the impacts of insecticides on them. There is evidence that neonicotinoids have been greatly impacting bees and other pollinators which is problematic for ecosystems and food security.

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?

```
Conc_class <- class(Neonics$Conc.1..Author.)
#finding the class of concentration, which is a factor
Conc_class
```

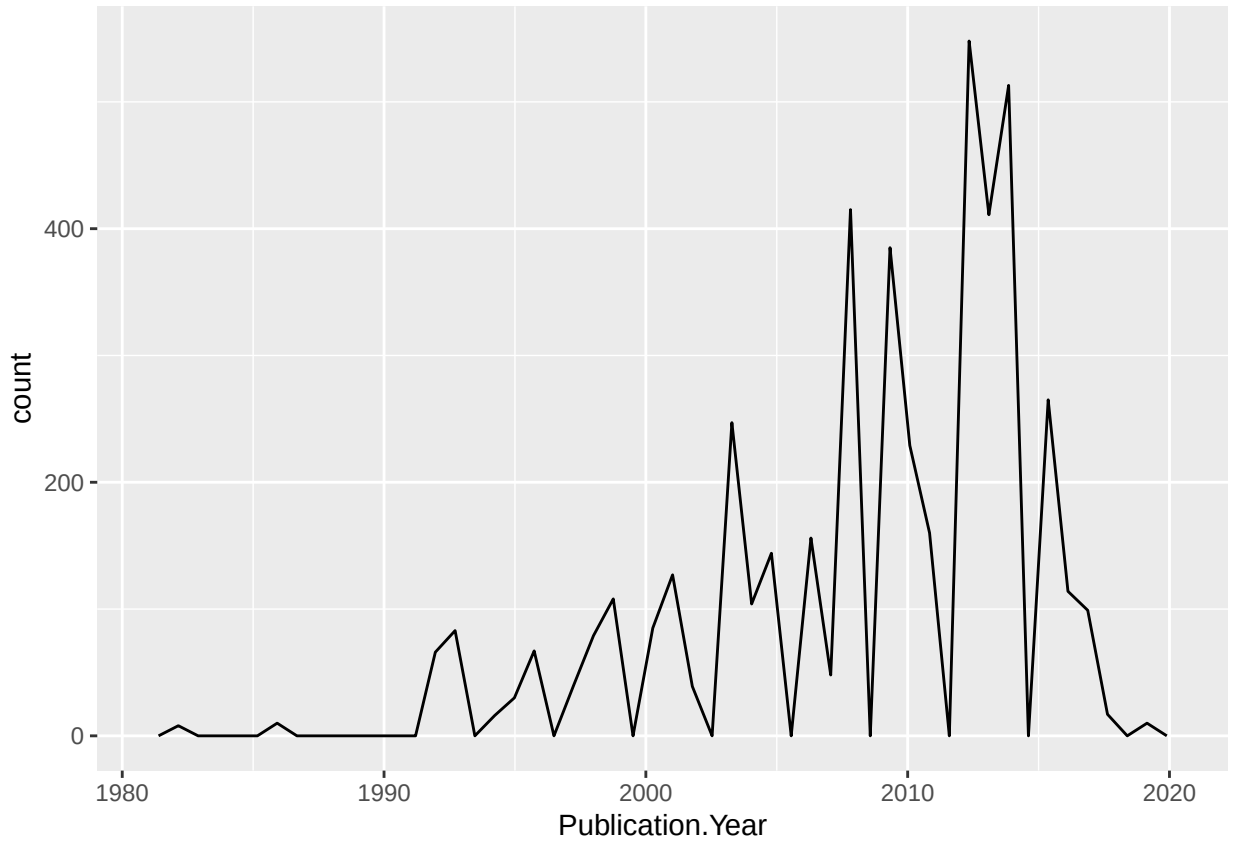
```
## [1] "factor"
```

Answer: It is a factor because some of the entries include non-numerical values such as slashes and 'NR'.

Explore your data graphically (Neonics)

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

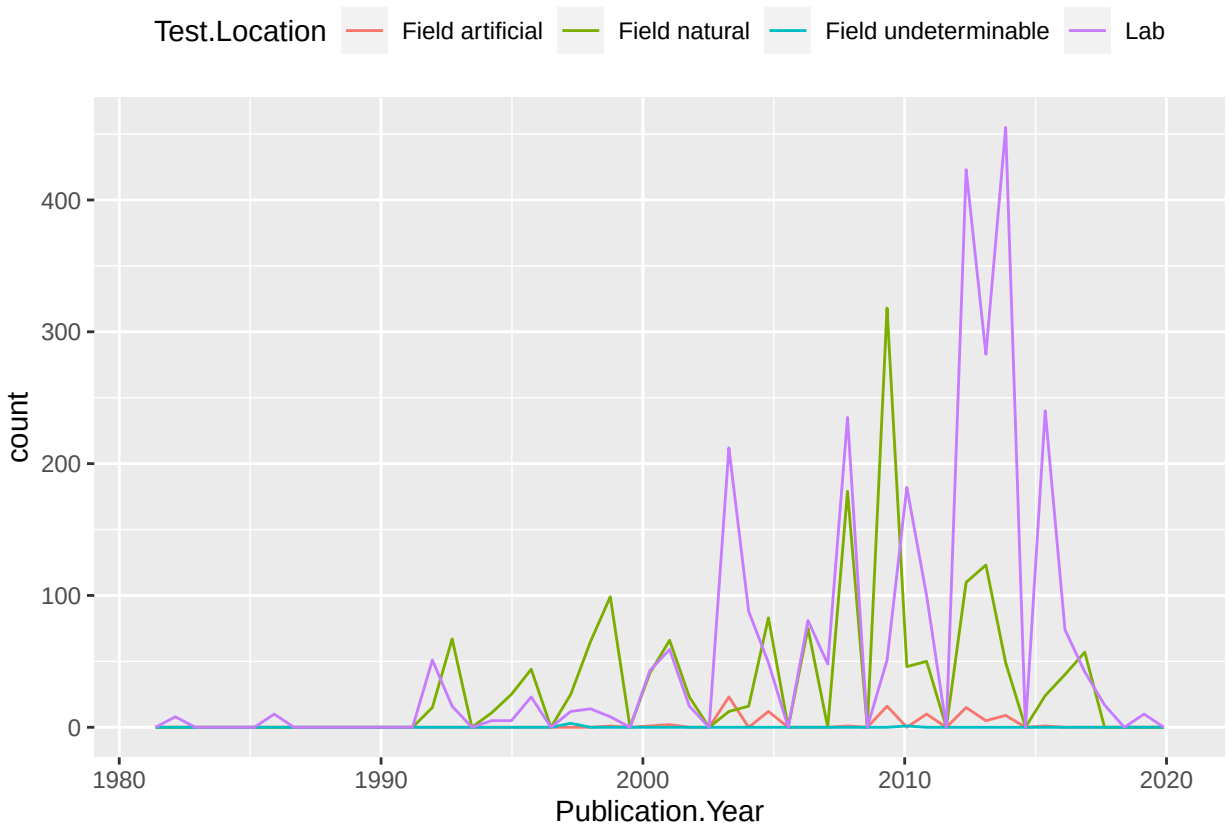
```
study_freq <- ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year), bins = 50)
study_freq
```




```
#frequency plot of the number of studies per year
```

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

```
study_freq2 <- ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year,
                                                   color = Test.Location),
                                               bins = 50) +
  theme(legend.position = "top")
study_freq2
```




```
#frequency plot of the number of studies per year separated by test location
```

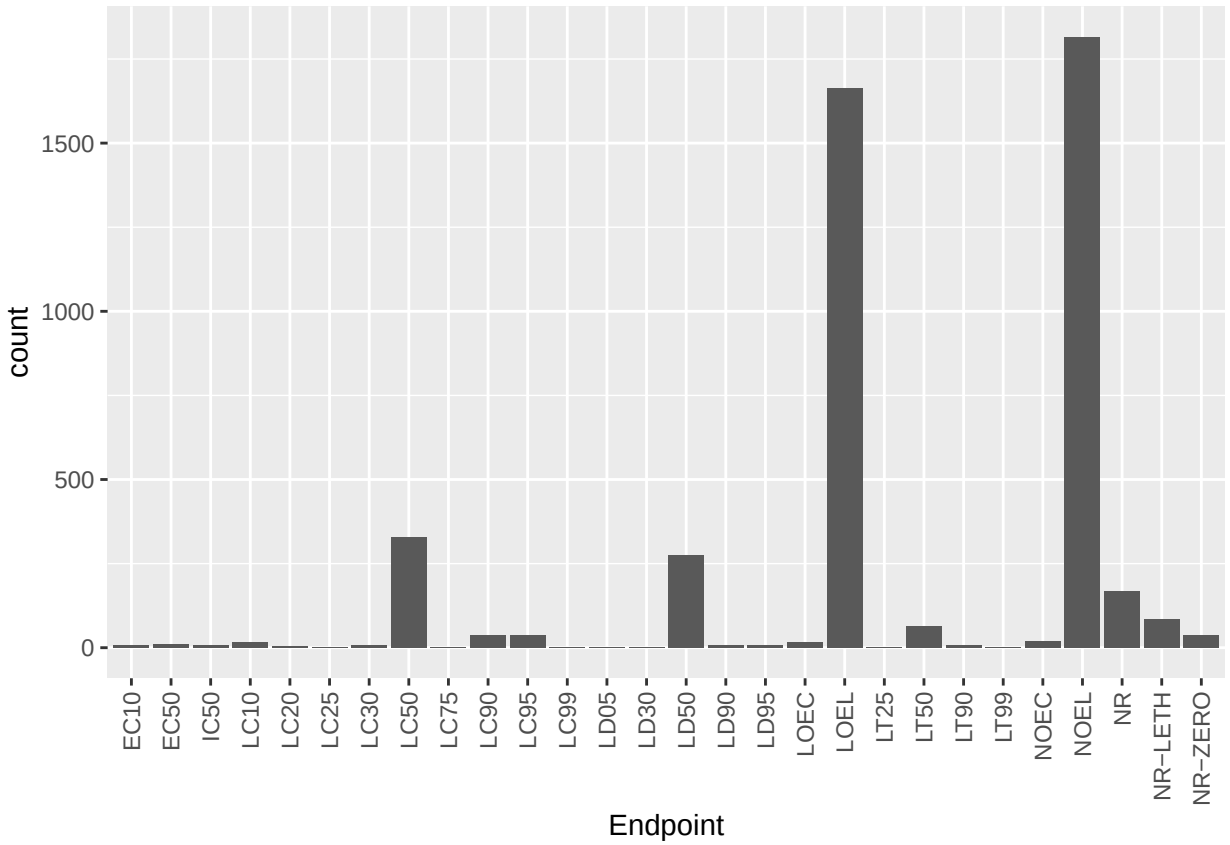
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are the lab and the natural field. These two are always the most common test locations, but there is variation over time, with the natural field being more common between 1990 and 2000 and the lab becoming more common overall after about 2002. The frequency of lab studies has a large increase between 2010 and 2015.

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...]

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




```
#Bar plot displaying the number of each endpoint included in the dataset
```

Answer: The two most common endpoints are NOEL and LOEL. NOEL is the 'No observable effect concentration' which represents the highest concnetration at which there is no adverse effect observed. The LOEL is the "Lowest observable effect concentration' and is defined as the lowest concentration at which an adverse effect is seen in the organism.

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.

```
Collect_class <- class(Litter$collectDate) #checking the class
Collect_class
```

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

```
#using lubridate to change the class
Litter$collectDate <- ymd(Litter$collectDate)

#checking the class again
Collect_class2 <- class(Litter$collectDate)
Collect_class2
```

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

```
#looking at the sampling dates
SampleDates <- unique(x = Litter$collectDate, incomparables = F)
SampleDates #Aug 2 and Aug 30 were sampled
```

```
## [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`?

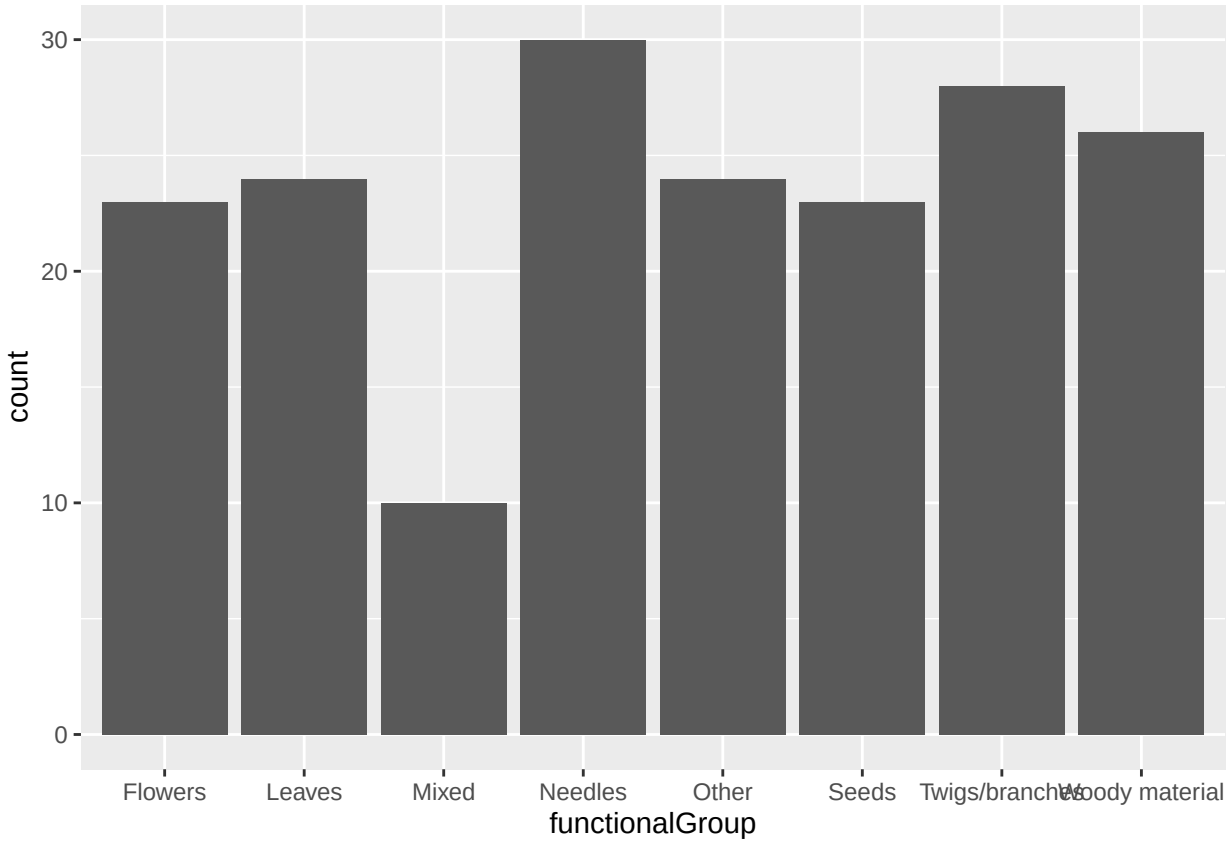
```
Plots <- unique(Litter$plotID, incomparables = F) #looking at the unique plots
Plots #12 different plots
```

```
##  [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
```

Answer: The unique function returns one of each unique entry to show how many different entries there are, while the summary function returns the count for each unique entry in that column.

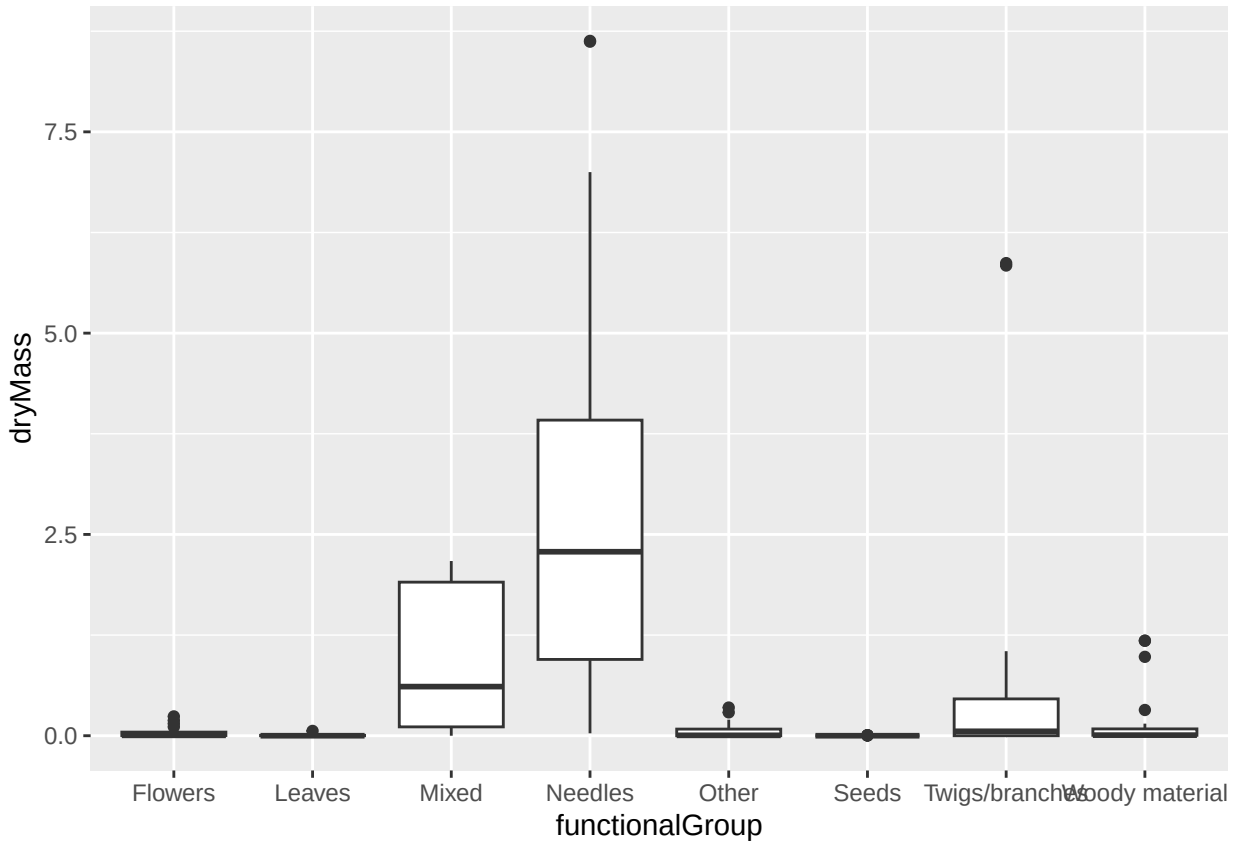
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
Litt_type <- ggplot(Litter, aes(x = functionalGroup)) + geom_bar() #bar plot
Litt_type
```

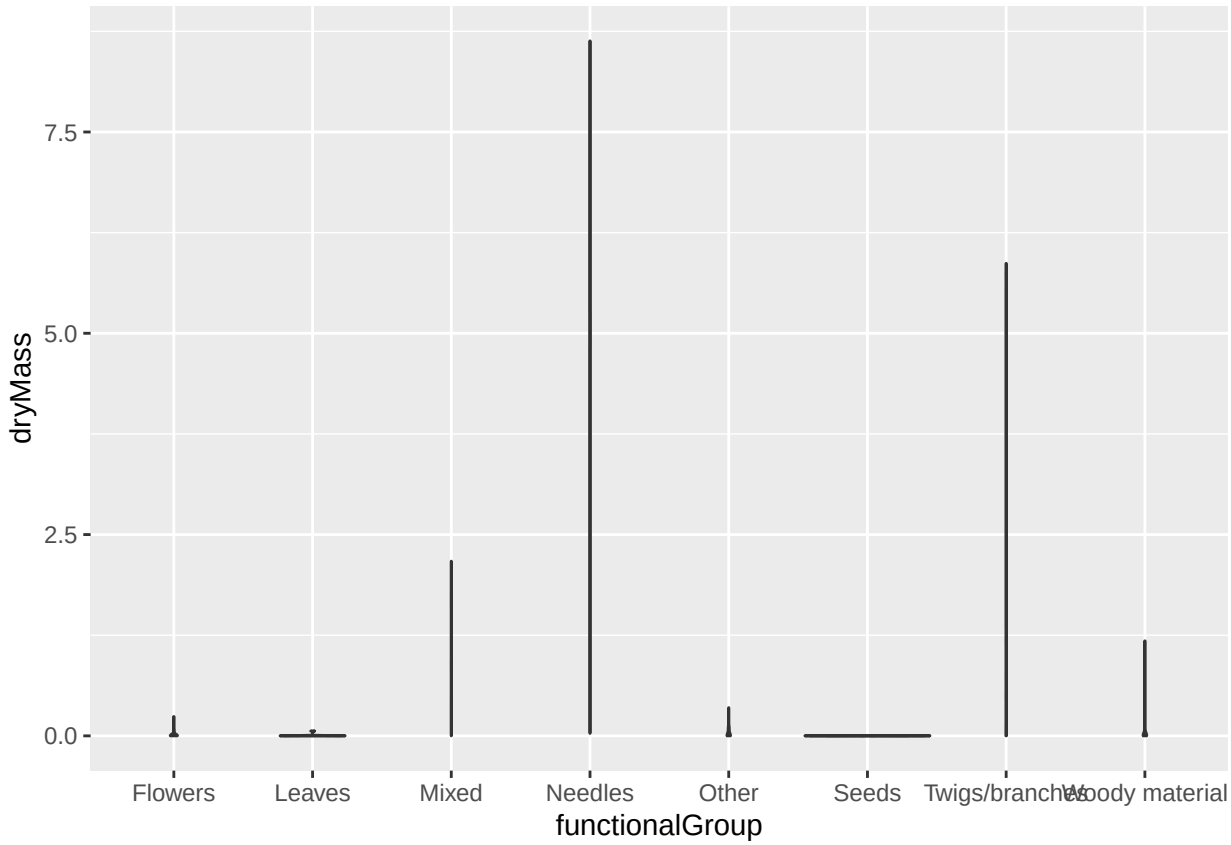


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
mass_box <- ggplot(Litter) + geom_boxplot(aes(x = functionalGroup,
                                              y = dryMass))
mass_box #boxplot of dry mass for each functional group
```




```
mass_violin <- ggplot(Litter) + geom_violin(aes(x = functionalGroup,
                                                y = dryMass),
                                            draw_quantiles = c(0.25, 0.5, 0.75))
mass_violin #violin plot of dry mass for each functional group
```



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

> Answer: The boxplot is more effective because functional group is a discrete variable and therefore has no range to be shown in the violin plot. A sample can only be a functional group or not be that functional group, so there is no variation in the x-axis for this set of data and the violin plot therefore only shows straight lines. In the boxplot, the spread of the data and the quantiles are visible.

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

> Answer: At these sites, the mixed and needles functional groups tend to have the highest biomass.