Assignment 3: Data Exploration

Erin Ansbro

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() #making sure my Working Directory is pointed to my Class Folder
```

```
## [1] "C:/Users/eka19/OneDrive - Duke University/Documents/872/EDE_Fall2023"
```

```
#install.packages("tidyverse") #installing tidyverse, removing for knit
library(tidyverse)
#install.packages("lubridate") #installing lubridate, removing for knit
library(lubridate)
Neonics<-read.csv("./Assignments/Week03_Assignment/ECOTOX_Neonicotinoids_Insects_raw.csv",
```

```
                  stringsAsFactors = TRUE)
#importing data sets and renaming them

Litter<-read.csv("./Assignments/Week03_Assignment/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: Knowing the ecotoxiocology on insects could help us in preventing crops being eaten or destroyed by insects. It's very helpful to know what insecticides work and on what insects. In addition, insecticides could cause issues to other animals and humans, and it would be helpful to have a data set of what insecticides are used widely in agriculture for that purpose.

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 might be interested because of the cycle of energy. Plants can use fallen leaves and debris as nutrients that can help fertilize the trees, to make them grow stronger, and ultimately better the health of the forest. Knowing the amount of debris could then help humans determine how a forest might be in terms of its helath.

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. Spatial sampling design uses tower plots to collect data.
   2. Temporal sampling design uses ground traps and elevated traps. Ground traps are checked once a year, elevated traps are checked based on type of forest - decidous can be checked up to twice a month, and evergreens can be once every one to two months.
   3. Trap placement within plots for spatial sampling may be either random or targeted, depending on the vegetation.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) #finding dimensions for Neonics data set, it is 4623 rows by 30 columns
```

```
## [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?

```
summary(Neonics$Effect) #finding common effects for Neonics data set
```

```
##     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: Population and Mortality are the the most popular at 1803 and 1493 respectively. Since this data set is about insecticide, it would be helpful to know what the inseciticide does to the populaton and mortality of insects to get a sense of how effective the insecticide is.

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. . .]

```
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 Wooly 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
```

```
#finding more info about Neonics, specifically species common name.
#Also using sort to find the most common.
```

Answer: (Other) 670, Honey Bee 667, Parasitic Wasp 285, Buff tailed Bumblee 183, Carniolan Honey Bee 152, and Bumble Bee 140. These species are mostly bees and other pollinators. If the insecticide affects pollinators, that could have large ramifications for fertilization and reproduction of plants.

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.) #Finding class of authors column in Neonics
```

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

Answer: The class is factor. It is not numeric becuase it looks like different people used different units, and types of concentrations. There is not a uniform way of measuring the concentrations. There are more than numbers, sometimes there is a / or an NR. Maybe for this reason, it is a factor, to categorize and to make it easier to understand that there is not uniformity within this column and it is more about understanding how the concentration was made instead of the actual concentration amount.

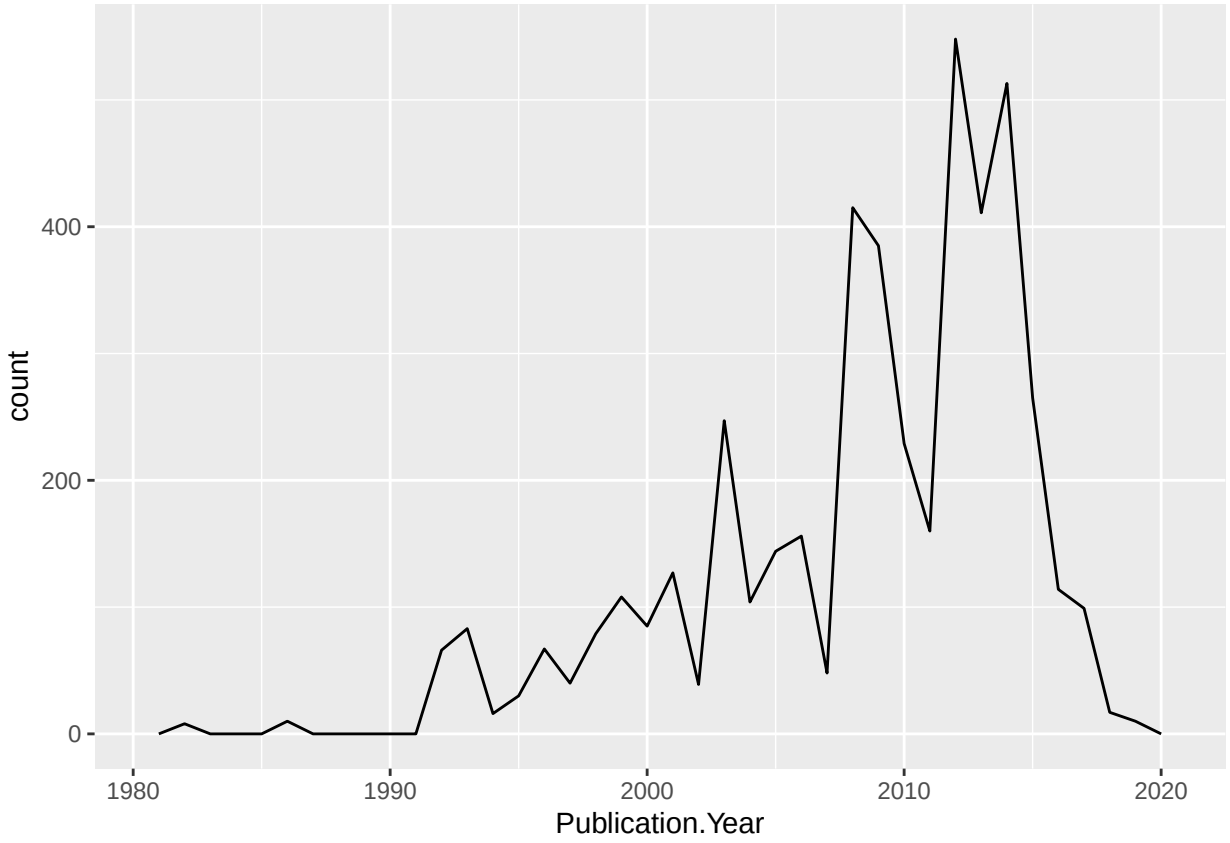
Explore your data graphically (Neonics)

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

```
sort(summary(Neonics$Publication.Year))
```

```
##     Min.  1st Qu.     Mean   Median  3rd Qu.     Max.
## 1982.000 2005.000 2008.431 2010.000 2013.000 2019.000
```

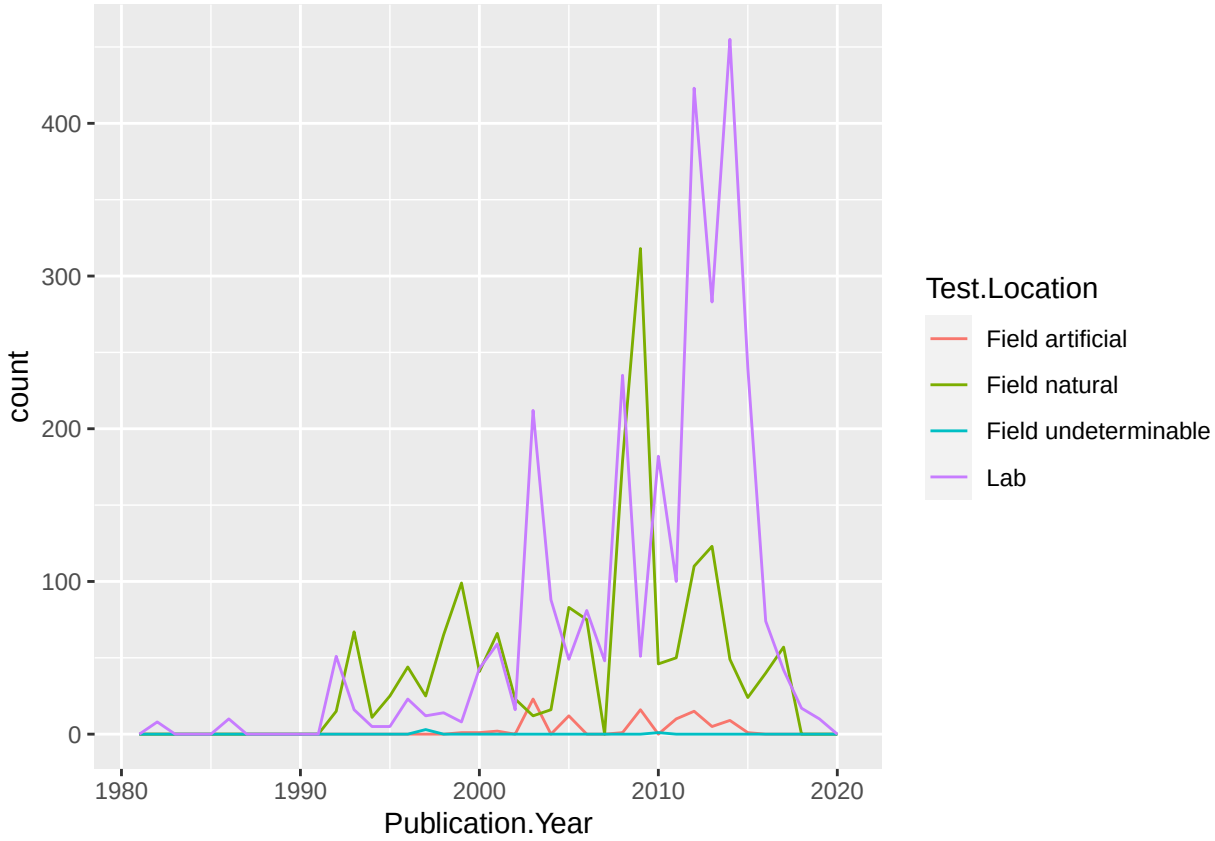
```
#lowest date is 1982, highest is 2019, which means there are 38 bins
library(ggplot2)
ggplot(Neonics) +
  geom_freqpoly(aes(x=Publication.Year), bins = 38)
```




```
#creating plot for studies per publication year
```

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

```
#this time with color
ggplot(Neonics) +
  geom_freqpoly(aes(x=Publication.Year, color = Test.Location), bins = 38)
```



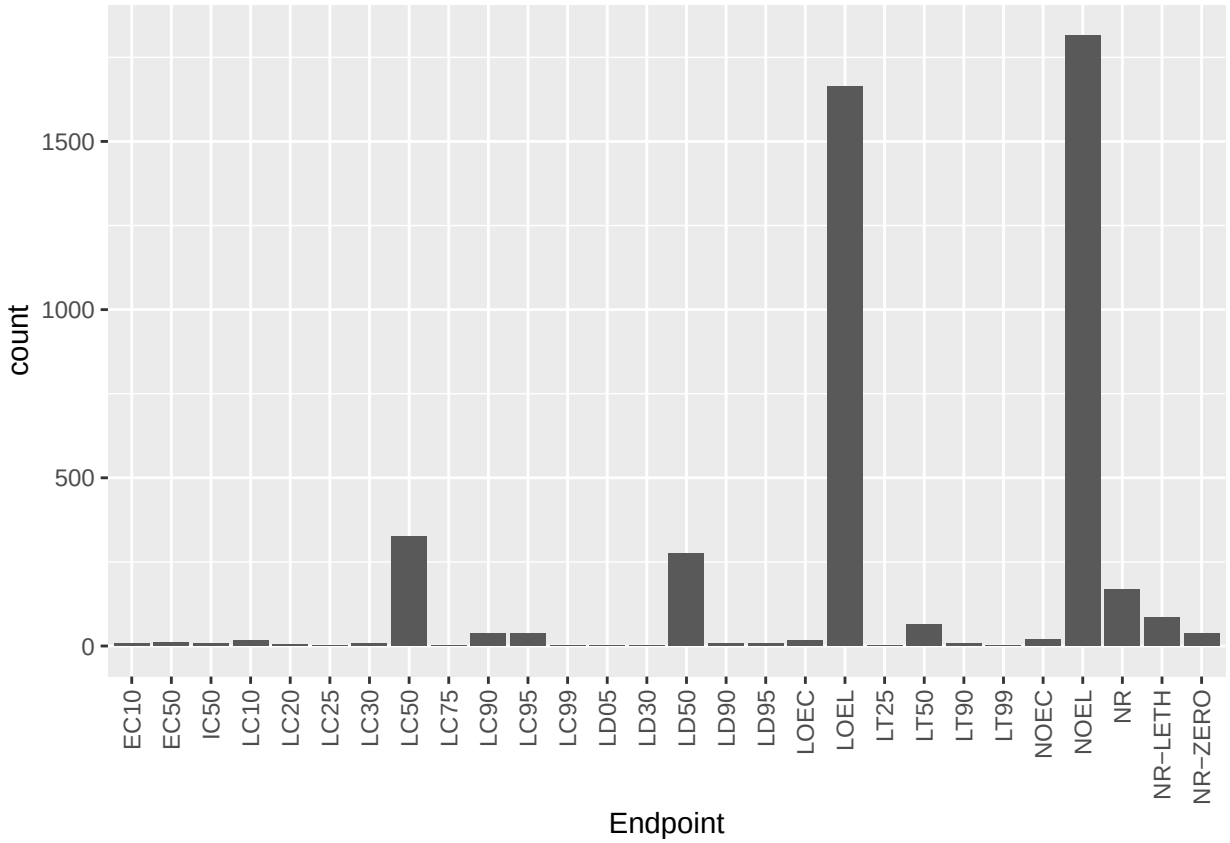
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: Lab is most common, then field natural, field artificial, field indeterminate. Overtime, the amount of test locations in the lab increased. There was a peak of field natural in the latter half of the 2000s (around 2005-2010), but has remained fairly steady over time besides that.

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))
```




```
#finding the two most common endpoints
```

Answer: NOEL and LOEL. Noel defined as No observable effect level which means the highest dose that produced effects was not significantly different from the results of the control.LOEL is lowest observable effect level which means the lowest dose that produced effects that were significantly different from control results.

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.

```
class(Litter$collectDate) #finding out if collect date is a date
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate) #changing to date
class(Litter$collectDate) #seeing if changing to date worked
```

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

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

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

```r
#finding unique collect dates, they are 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`?

```r
unique(Litter$plotID)#finding how many plots were sampled at Niwot Ridge
```

```
##  [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
```

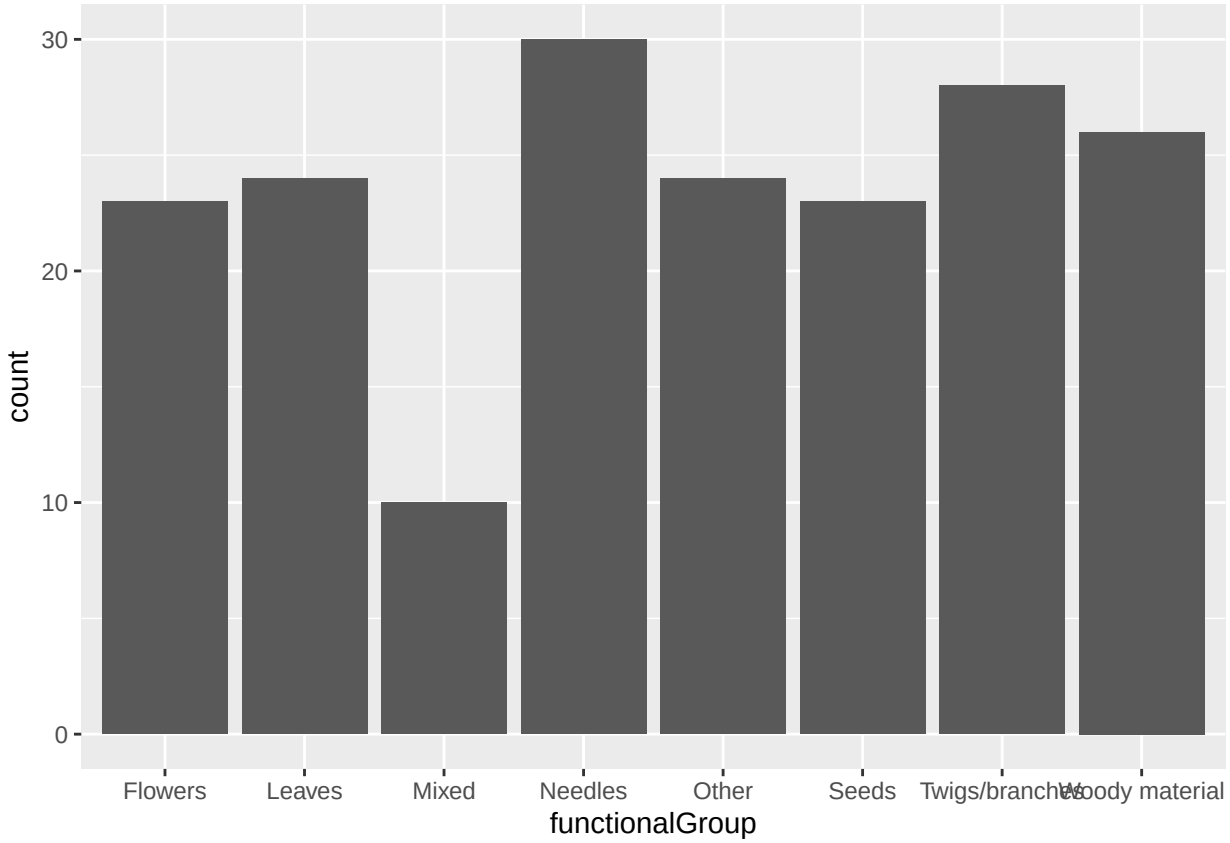
```r
summary(Litter$plotID)
```

```
## 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: Summary gives you an overview of the information in that column and the number of times a plot id is present, unique gives you the unique values for that column. In this case there are 12 plots. Summary gives you the number of times a specific plot is listed, while unique just provides it once.

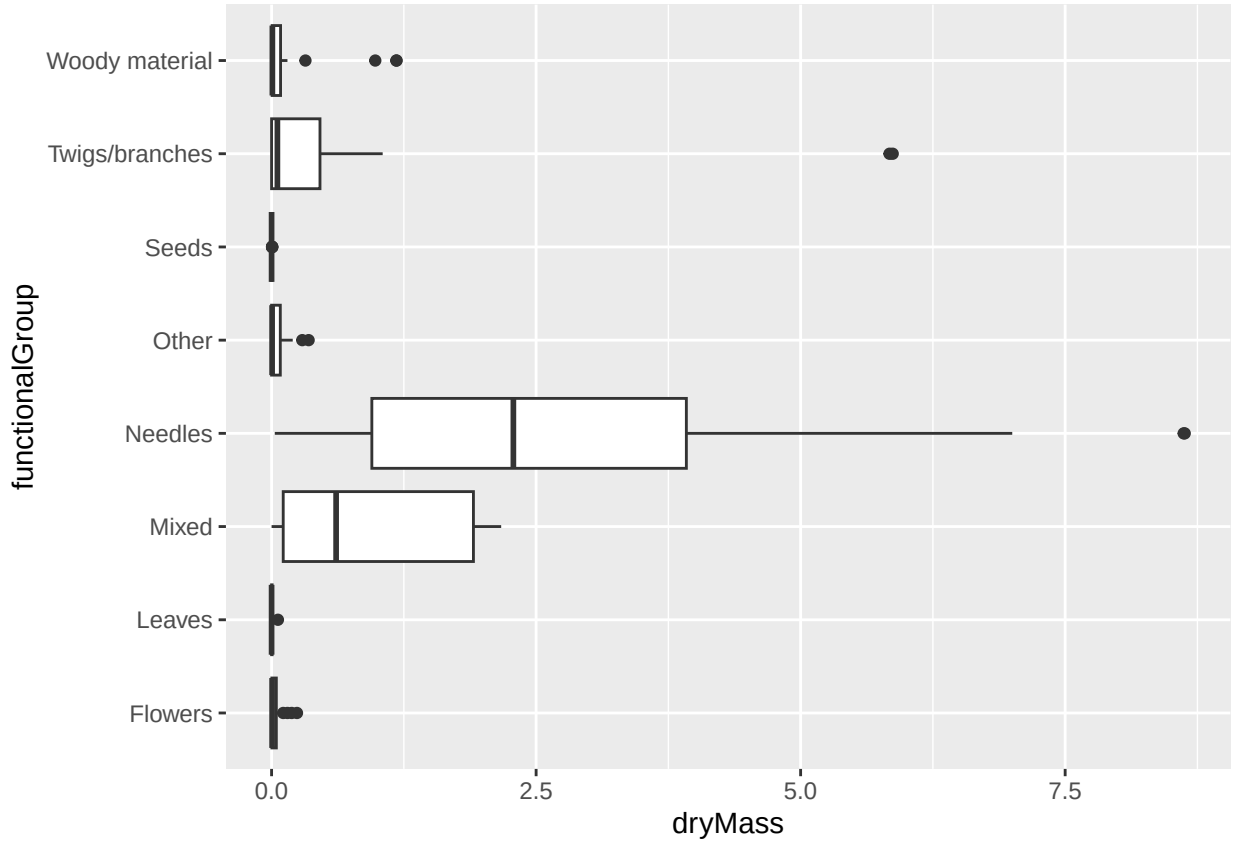
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
library(ggplot2)#calling ggplot
ggplot(Litter, aes(x=functionalGroup)) +
  geom_bar() #finding what type of litter is collected using bar plot
```



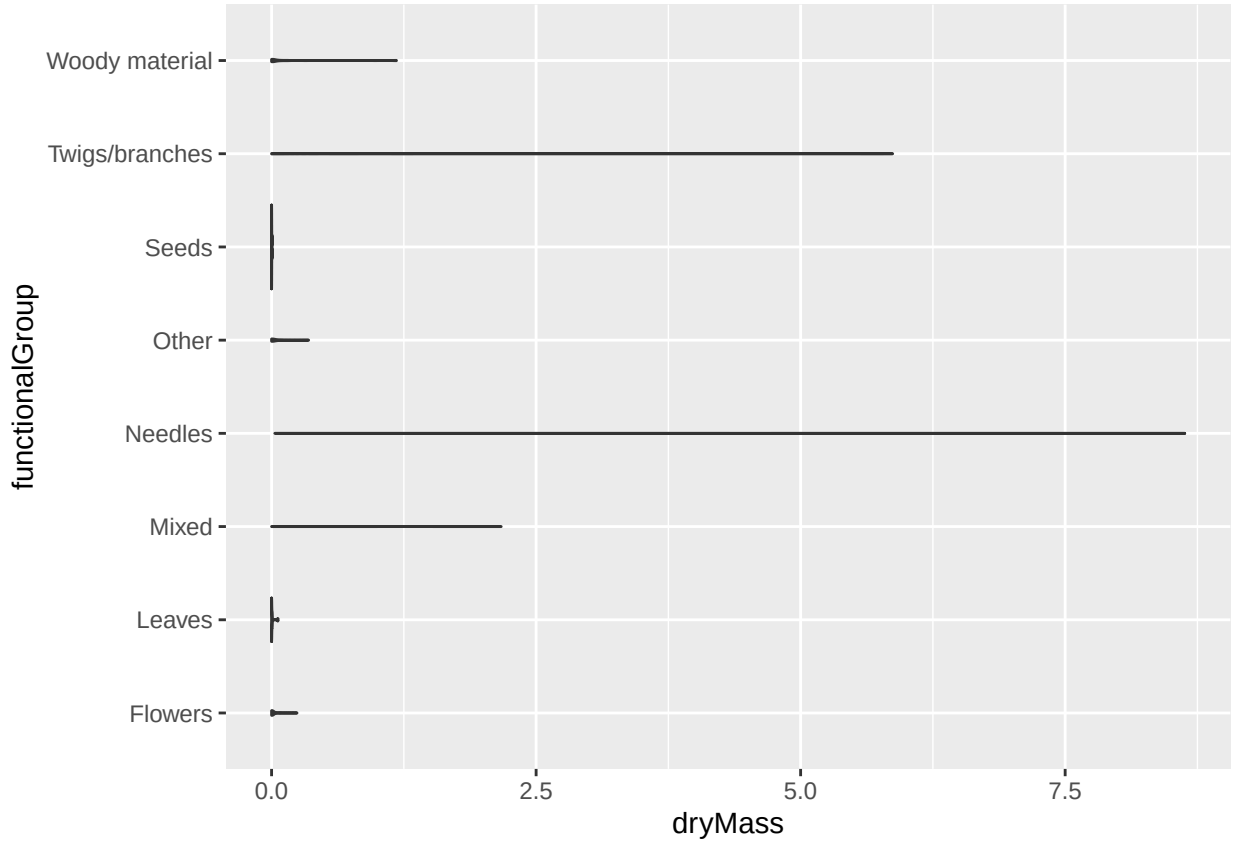
15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




```
#boxplot to compare with violin plot

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




```
#violinplot
```

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

> Answer: There are not a lot of results, making the violin plot appear thinner. Boxplot shows more summary, like interquartile range and median. This can be created from small and large data sets, so it is a better visual for a data set that is not too dense.

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

> Answer: Needles, then Mixed