Assignment 3: Data Exploration

Maia Griffith

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 directory
getwd()
```

```
## [1] "/home/guest/ENV872/maia-g/EDE_Fall2023"
```

```
#loading packages
library(tidyverse)
library(lubridate)

#Import datasets
```

```r
###ISSUE: using the './Data/' then tab to see the options yields nothing, but it was working when I had

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)

###QUESTION: Do I still need to call these out since it is enormous?
#Neonics
#Litter
```

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: Insecticides often kill all kinds of insects in an area, including beneficial ones, and can also harm other organisms that feed on the crop or the insects in that area. Pollinators like bees and butterflies for example are getting destroyed by this type of insecticide. Organisms like birds and fish are often one of the most impacted by insecticide use, especially since neonics stay in the environmental system (soil, water, fruit, etc) for many years.

   Additional background info found here: https://www.nrdc.org/stories/neonicotinoids-101-effects-humans-and-bees

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 woody debris in forests tells foresters about the rate of decay within that ecosystem as well as provides insights into fire likelihood and how big a fire would be. More debris, especially dry debris, means a higher liklihood of fires. Debris also helps calculate Aboveground Net Primary Productivity, biomass estimates, and carbon fluxes according to the NEON Litterfal User Guide.

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. One elevated litter trap and one ground trap deployed per 400m^2 plot 2. Ground traps sampled once per year, elevated traps depends on vegetation type 3. Trap placement is randomized in places with 50% > cover and targeted in areas with less vegetation.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
dim(Neonics)
```

```
## [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?

```r
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: Population (1803) and Mortality (1493) are by far the most common effects studied, probably because it is really important for biologists to know how big an insect population is in an area and how many die, especially when looking at impacts of an insecticide.

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...]

```r
#Created a variable that will sort the common name data
sorted_common_name <- sort(summary(Neonics$Species.Common.Name))
sorted_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
```

```r
#Then created a variable to hold and show me the top 6 most commonly studied species
top_6 <- sorted_common_name[95:100]
top_6
```

```
##            Bumble Bee   Carniolan Honey Bee Buff Tailed Bumblebee
##                   140                   152                   183
##        Parasitic Wasp             Honey Bee               (Other)
##                   285                   667                   670
```

Answer: While the largest category is `Other`, the others in the top 6 are all pollinators. If we ignore the `Other` category and include the next most common species, then we still get all bees and one wasp. As key pollinators, these organisms help maintain healthy ecosystems and are therefore of high interest for conservation, especially since they are one of the most negatively impacted groups as I mentioned in Question 2.

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?

```r
class(Neonics$Conc.1..Author.)
```

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

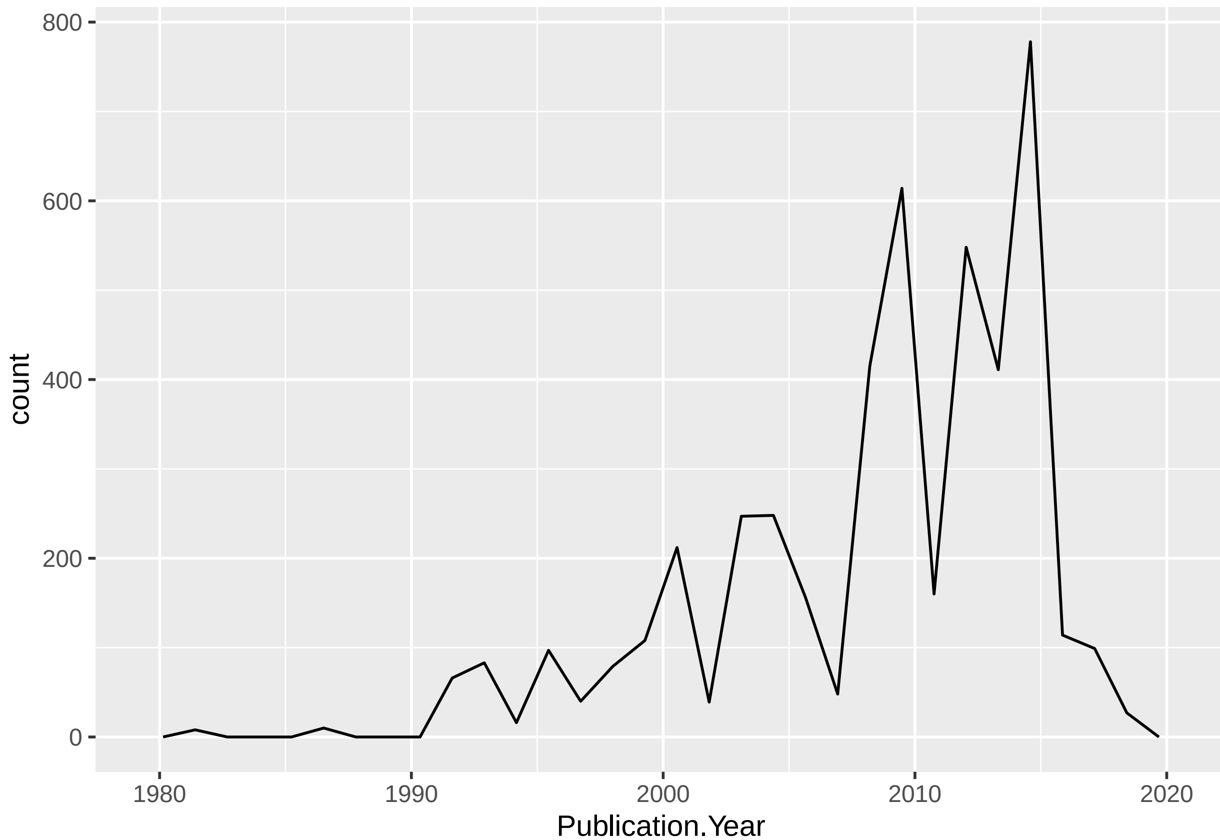
Answer: The class is "factor" because when we imported the dataset we made all the data into factors using the `stringAsFactors = T` command.

Explore your data graphically (Neonics)

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

```
###QUESTION: Do I need to somehow call out each plot so it shows when I knit or does it automatically s
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year))
```

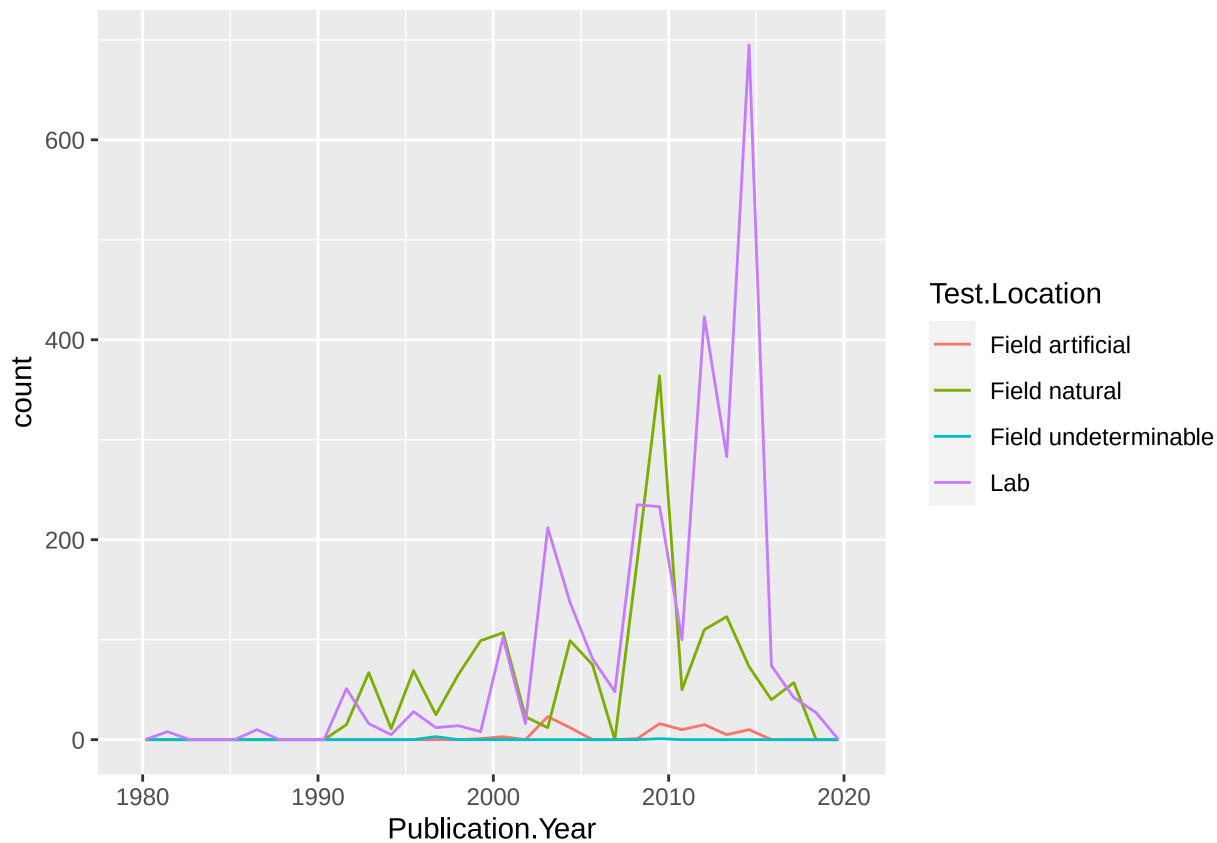
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

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

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



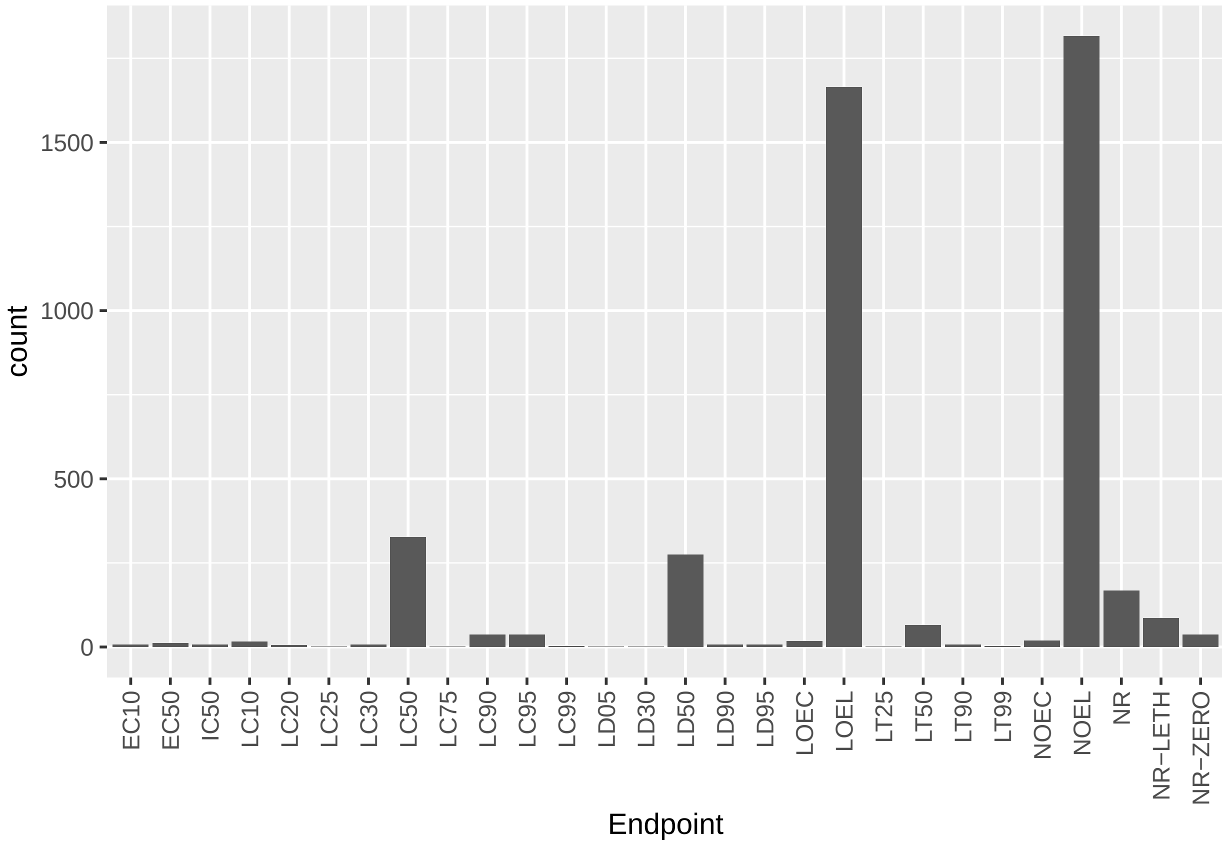
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are Lab and Field natural, although Lab definitley increases between 2010 and 2020, perhaps because field work is generally more expensive than lab work. The overall number of studies increased as time went on as well.

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 LOEL and NOEL, which are both terrestrial-based codes. LOEL is defined as the "Lowest-observable-effect-level" in the Code Appendix, meaning the lowest concentration of toxins that resulted in significant effects. NOEL is defined as "No-observable-effect-level" in the Code Appendix, meaning the highest concentration of toxins that DO NOT result in significant effects.

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) #The class is factor.
```

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

```
#Determining what format the dates are currently written in.
Litter$collectDate[1] #This shows it is in year-month-day format
```

```
## [1] 2018-08-02
## Levels: 2018-08-02 2018-08-30
```

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

#Check to see if it changed.
class(Litter$collectDate) #this vector is now a Date class.
```

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

```r
#Using the `unique` function.
unique(Litter$collectDate) #shows Aug 2nd and 30th were sample dates.
```

```
## [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`?

```r
#Use the length AND unique functions together to return the number of plots sampled.
length(unique(Litter$plotID))
```

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

```r
#Comparing `summary` and `unique` functions.
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
```

```r
unique(Litter$plotID)
```

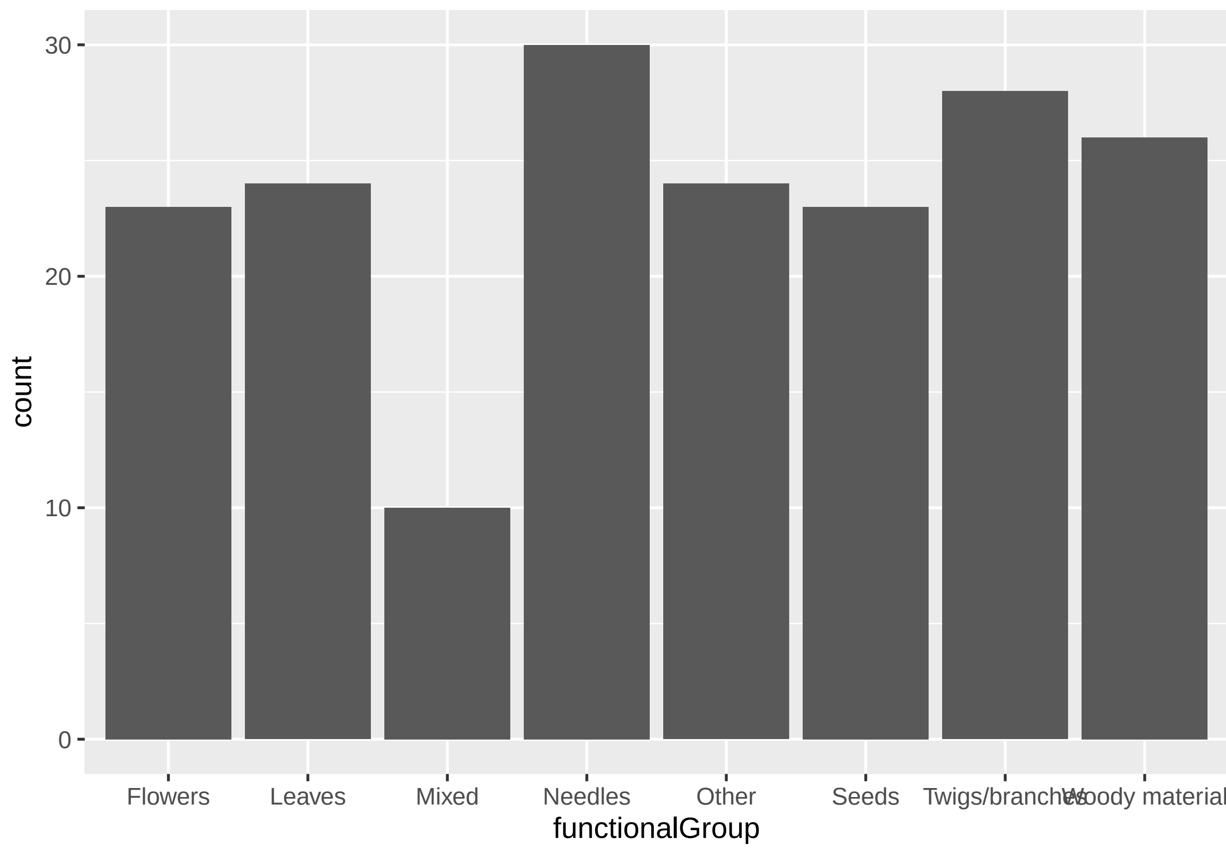
```
##  [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

```r
###QUESTION: What are Levels? Google was not helpful...
```

> Answer: Summary returns the plot IDs AND how many times each was sampled. Unique returns all the different plot IDs and excludes repeats. It also includes "Levels" which shows the plot IDs in ascending order.

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.

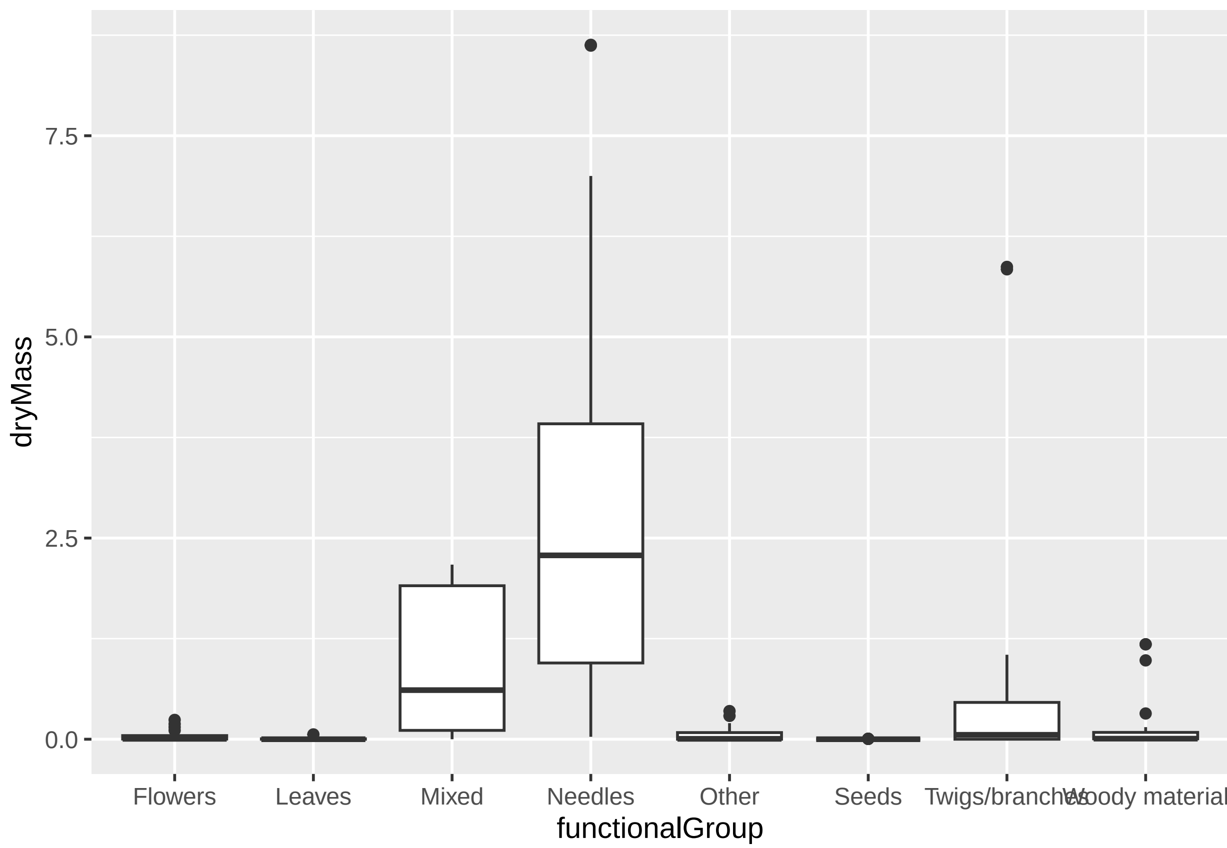
```r
#Used ggplot and geom_bar to make the graph.
ggplot(Litter) +
  geom_bar(aes(x = functionalGroup))
```



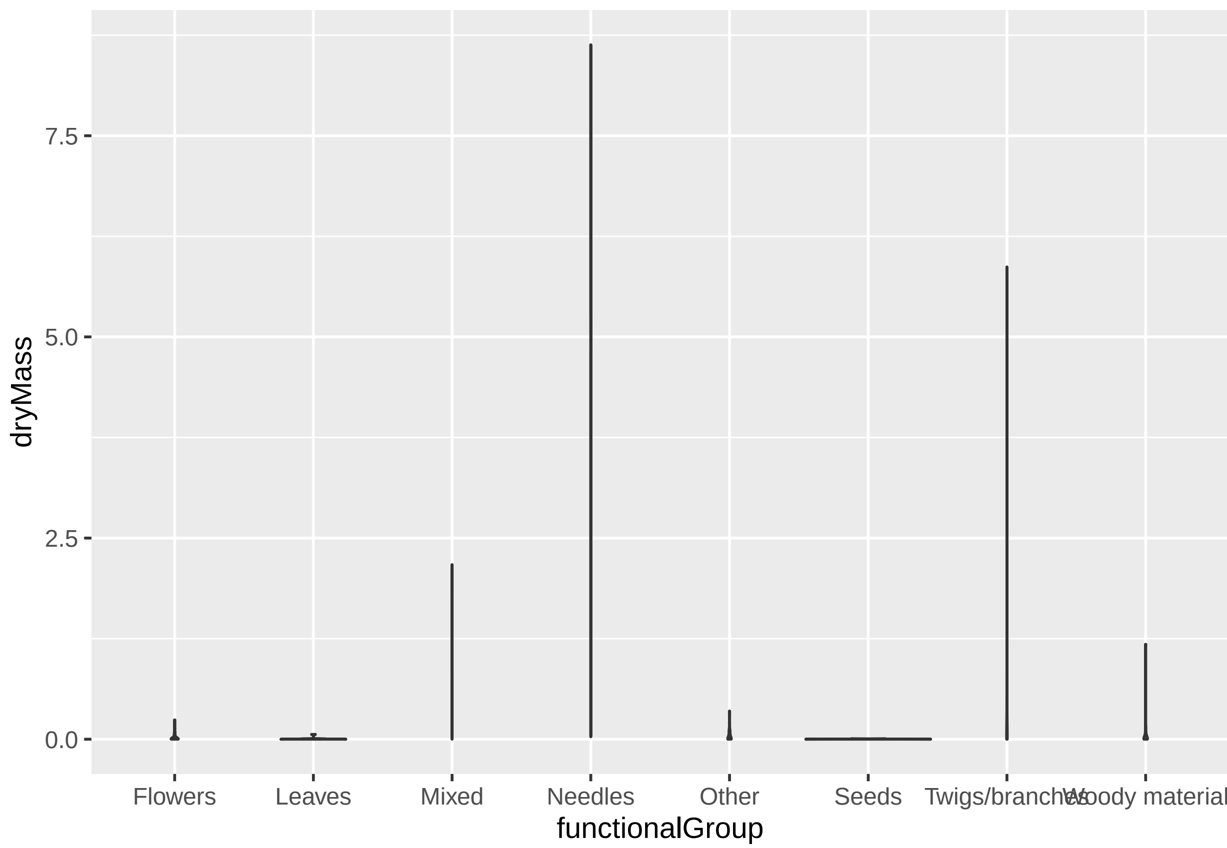

```
#the Functional Group "Mixed" has the least by quite a lot, but the others are fairly equally distribut
```

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

```
#First going to make the box plot with x-axis as functional group and y-axis as dryMass.
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass))
```




```r
#Next going to create a violin plot with same axis.
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass))
```




```r
#Violin plot does not look very good but I think I did it right.

length(Litter$dryMass) #Wanted to know how many data points we had
```

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

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

> Answer: The boxplot is more effective in this case because there are not a huge amount of data points to show any visible distribution. When we used the USGS dataset in the lesson, there were thousands of data points so the width of the violin plots were actaully discernable.

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

> Answer: Needles and Mixed tend to have the highest biomass, with one outlier in Twigs/branches that is a very high value.