Assignment 3: Data Exploration

Brynn

Fall 2024

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 Canvas.

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. Load necessary packages (tidyverse, lubridate, here), check your current working directory 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.

```
#install tidyverse, lubridate, and here
library(tidyverse)
library(lubridate)
library(here)
#check working directory
getwd()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```r
here()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```r
#load datasets
Neonics <- read.csv(
  file = here('Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv'),
  stringsAsFactors = TRUE
  )
Litter <- read.csv(
  file = here('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: Neonicotinoids are a pesticide that cannot target specific variants of insects. This can be extremely damaging to pollinators who will unsuspectingly land on a plant with neonicotinoids on it and soon after be struck by its toxic compounds. Obviously, we need pollinators who are keystone species and integral to biodiversity.

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: Woody debris can tell us a lot about forest health as well as the insects and other creatures that live there. Woody debris can be sampled for pathogens, species vulnerability (if one species is shedding more than others), as well as the creatures that live there. Leaf herbivory and wood boring insects commonly live in these woody debris and can be an indicator of populations and impact of their presence.

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. Occurs in randomly selected tower plots 2.In sites with forested tower airsheds, the sampling is targeted to take place in 20 40m x 40m plots. In sites with low-statured vegetation over the tower airsheds, sampling is targeted to take place in 4 40m x 40m tower plots 3. In sites with > 50% aerial cover of woody vegetation >2m in height, placement of litter traps is random

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#load summary and dimensions of Neonics
#summary(Neonics)
#dim(Neonics)
#commented out because would not knit
```

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? [Tip: The `sort()` command is useful for listing the values in order of magnitude...]

```
#shows the summary of the effect feature of the Neonics dataset
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 and then mortality are the most important effects. This would be important because seeing how certain populations are dwindling helps to better understand the effect of the pesticide. Mortality does generally the same thing. Both of these effects help understand what species may be in danger or have population numbers plummeting quickly.

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: Explore the help on the `summary()` function, in particular the `maxsum` argument...]

```
summary(Neonics$Species.Common.Name, maxsum=7)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee     Italian Honeybee
##                   152                   140                   113
##               (Other)
##                  3083
```

```
#used maxsum to show the top 7 results for this column because 6 compiled,
#all the other species into one category for the seixth variable
```

Answer: Honey Bee, Paraitic Wasp, Buff Tailed Bumblebee, Carnolian Honey Bee, Bumble Bee, Italian Honeybee. Almost all of these species are pollinators and pollinators are vital to our ecosystem. This shows how harmful this pesticide can be, especially in the broader scope of biological diversity and pollination.

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? [Tip: Viewing the dataframe may be helpful...]

```
summary(Neonics$Conc.1..Author.)
```

```
##     0.37/      10/      NR/       NR        1     1023    0.40/       2/
##       208      127      108       94       82       80       69       63
##        10   0.053/      100      50/     0.5/     0.03    0.05/     0.45
##        62       59       56       51       45       44       43       43
##      0.1/    0.45/     1.0/    2.27/       50    0.125     500/      0.5
##        42       40       40       40       36       33       33       32
##    0.048/    0.15/       1/       48    25.0/      12/    0.027      2.4
##        30       30       30       30       28       27       26       26
##      0.2/    0.56/     100/        3    0.01/    1000/       3/    0.336
##        25       24       23       23       22       22       22       21
##      1.5/     0.05      1.5    2.60/    20.0/        6    6.80/    62.5/
##        21       20       20       20       20       20       20       20
##     0.005     0.4/    0.18/     0.3/     1000       40 0.00355/      0.1
##        18       18       17       17       17       17       16       16
##       0.4     150/      300      80/    0.053     0.24     0.28     125/
##        16       16       16       16       15       15       15       15
##         9   0.0001  0.0004/   0.084/     0.15      0.6    12.5/   144.0/
##        15       14       14       14       14       14       14       14
##      350/    40.0/      48/       56      84/    0.17/      125       14
##        14       14       14       14       14       13       13       13
##        16       17   0.047/    0.25/    0.28/    1.28/    1.81/      112
##        13       13       12       12       12       12       12       12
##       150     2.5/       25      60/      75/    0.02/   0.025/     0.29
##        12       12       12       12       12       11       11       11
##     37.5/       4/        5  (Other)
##        11       11       11     1817
```

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

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

```
#this helps us see the concentration and summary of this column. It is a factor
```

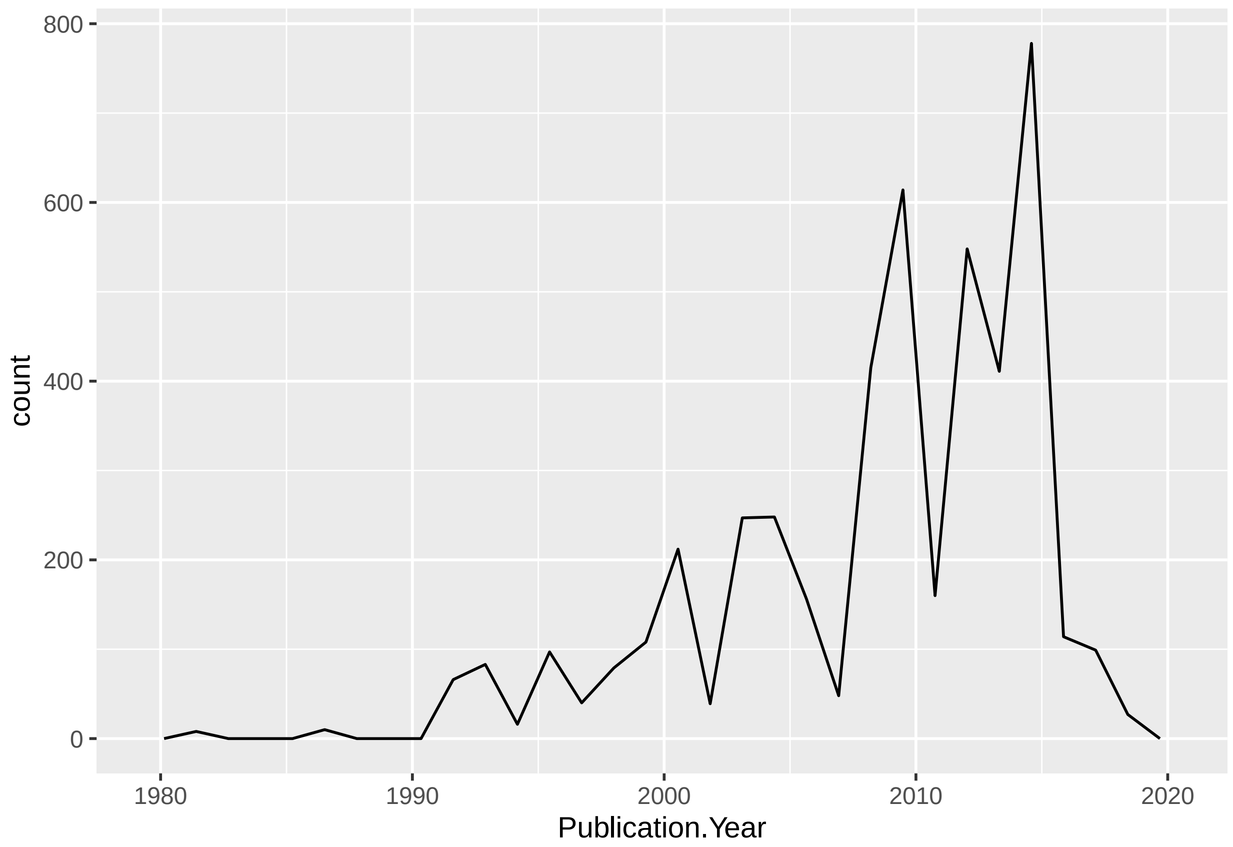
Answer: It is a factor, and there are non-number characters in column (/ for example), and it is not a full numeric answer either; it seems to be an unfinished calculation

Explore your data graphically (Neonics)

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

```
ggplot(Neonics) +
  geom_freqpoly(aes(x=Publication.Year))
```

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

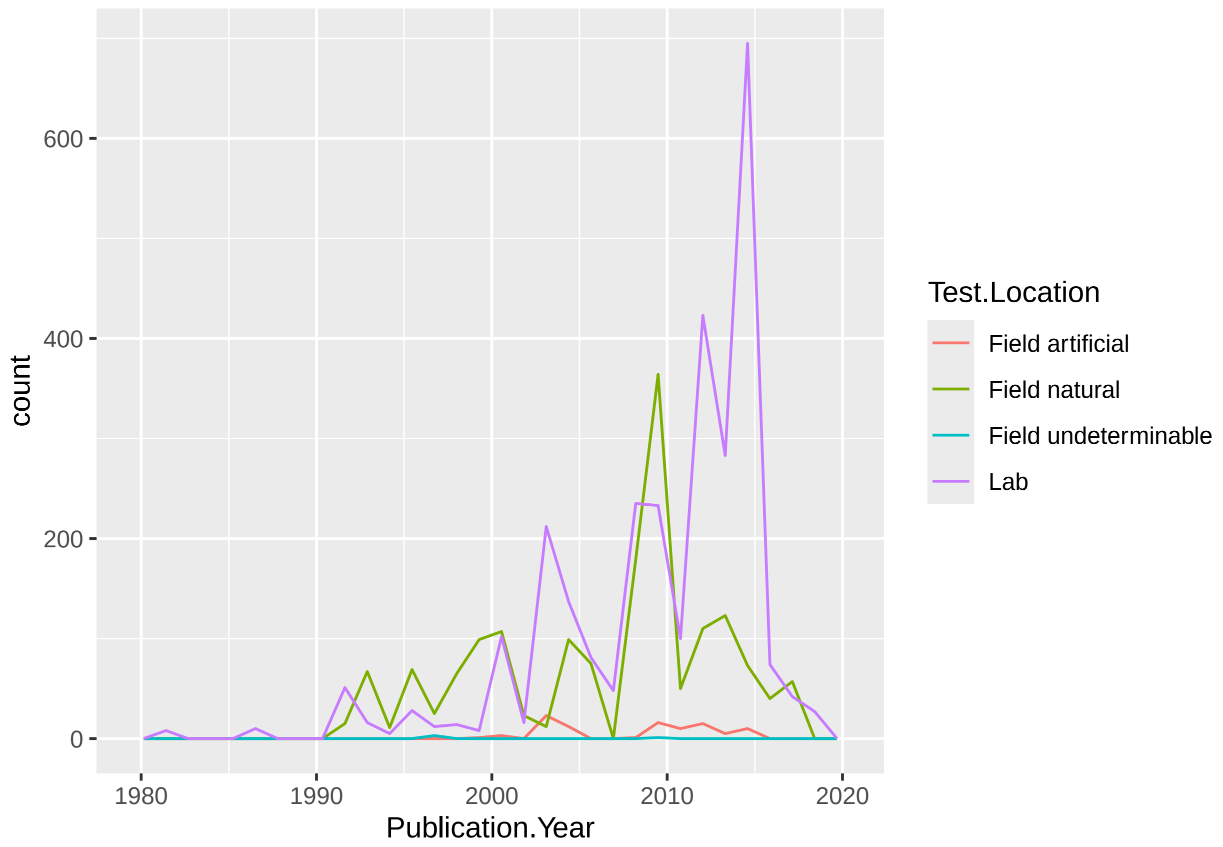


```
#uses ggplot and geom_freqpoly functions
```

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‘.
```



```
#keep color= inside of initial x factor and outside of aes
```

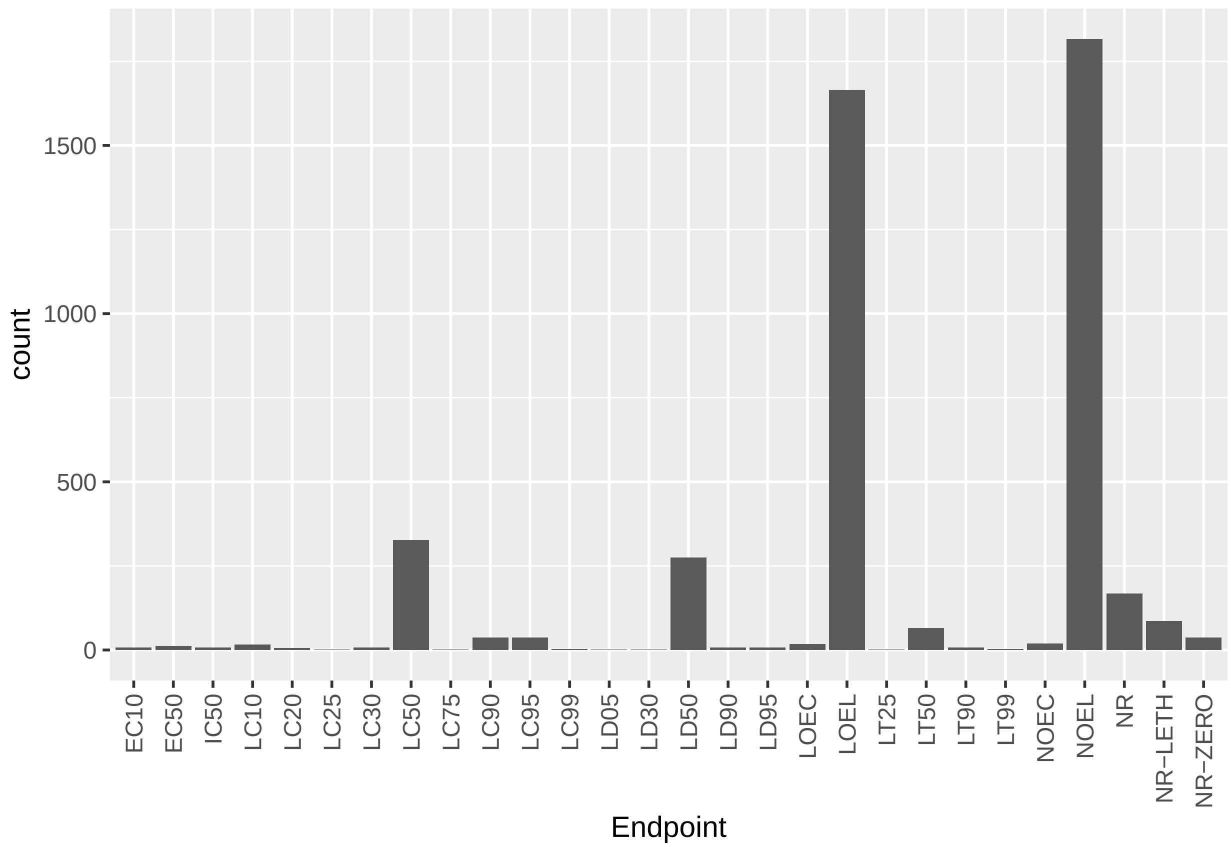
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are the lab which grew over time and in the field which peaks around 2019 but is otherwise somewhat insignificant. The graphs fall and rise which makes sense because certain insects may be more prevalent in some years than others due to brooding and hibernation years.

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))
```



```r
#ggplot and geom_bar were used here
```

> Answer: LOEL and NOEL are the most common endpoints in this data set. LOEL is the lowest observed effect level and the NOEL level means there was no observed effect. LOEL is the lowest dose of the toxicant that they were testing where an effect was observed, and conversely NOEL is the highest dose of the toxicant observed where there was no effect.

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.

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

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

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

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

```
#August 2nd and August 30th in 2018 were the dates litter was collected.,
#The unique function will show the repeated values only once wheras summary,
#whill show all of the values even if they are repeated.
```

13. Using the `unique` function, determine how many different plots were sampled at Niwot Ridge. How is the information obtained from `unique` different from that obtained from `summary`?

```
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
```

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

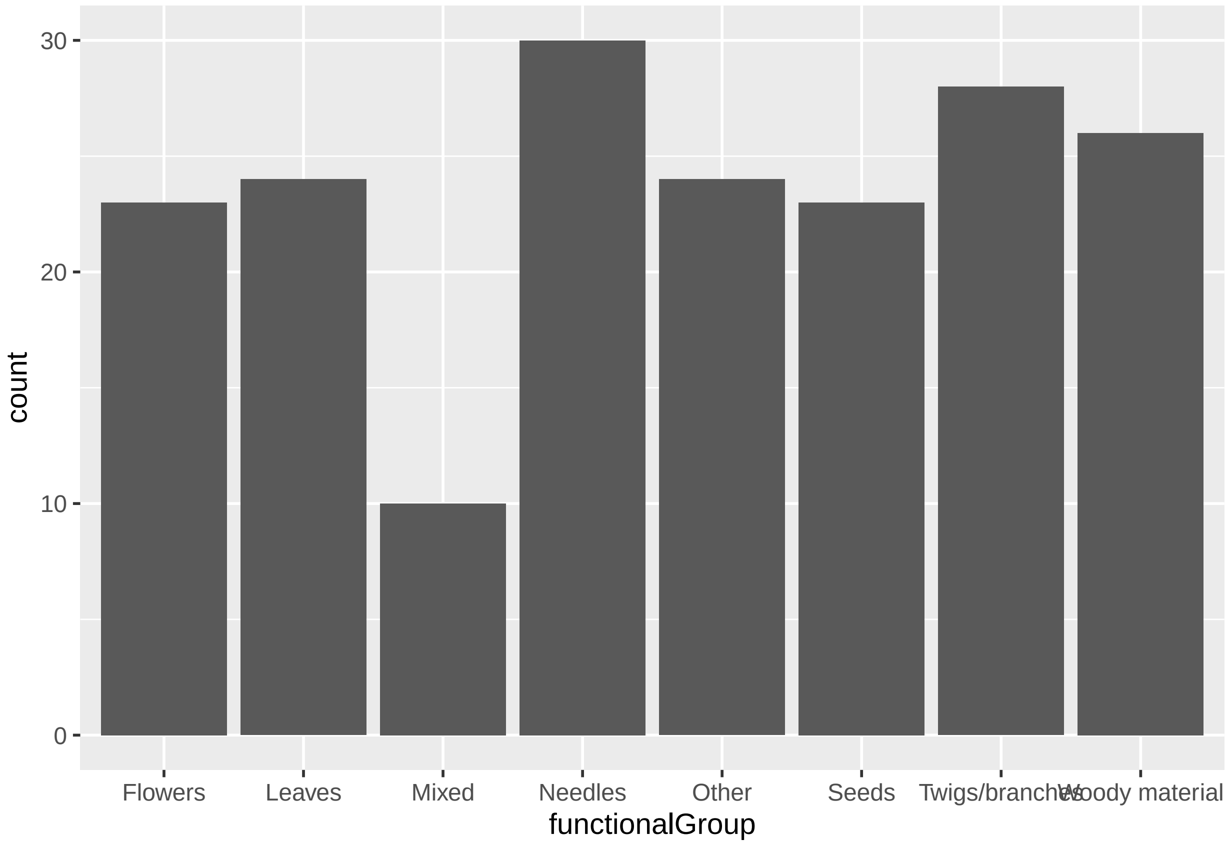
```
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061
##       20       19       18       15       14        8       16       17
## NIWO_062 NIWO_063 NIWO_064 NIWO_067
##       14       14       16       17
```

```
#shows the plot ID of both the unique and summary stats of litter sampling
```

Answer: The unique function returns the plot ID's, whereas the summary function returns, both the plot ID and the count for each plot ID as well.

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.

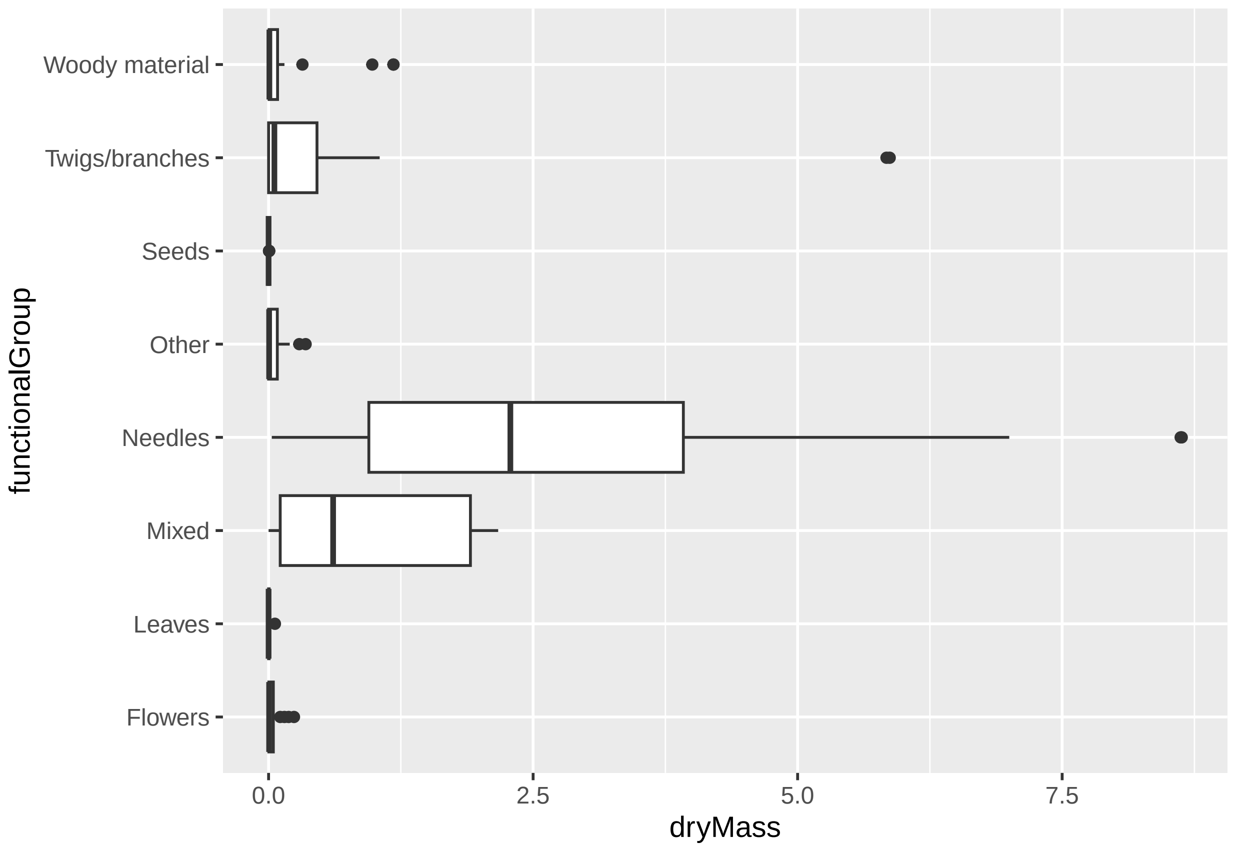
```
ggplot(data = Litter, aes(x = functionalGroup)) +
  geom_bar()
```



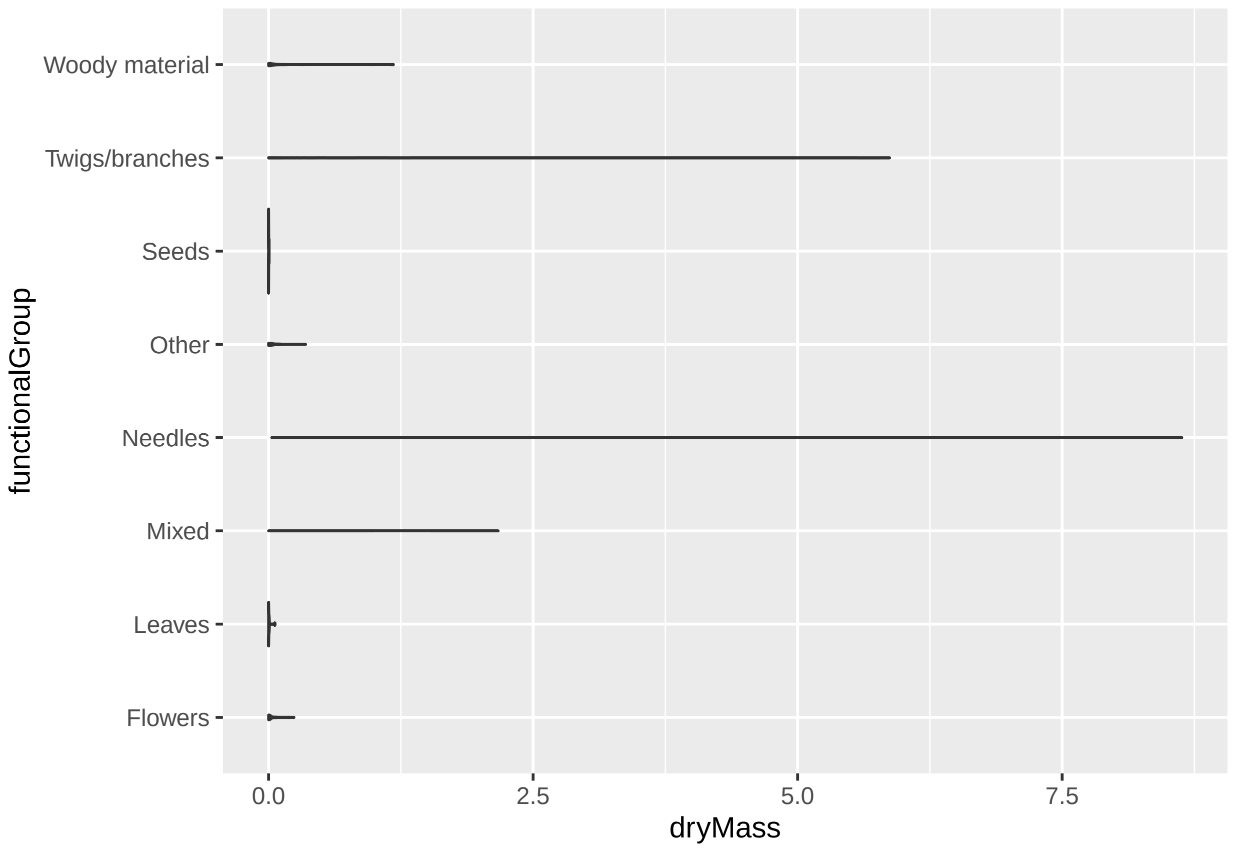
```
#geom_bar for bar graphs
```

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))
```



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



```
#geom_boxplot and geom_violin for these
```

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

> Answer: The boxplot shows the outliers and is easer to visualise. It also shows the average and standard deviation of each Dry Mass category. The violin on the other hand is hard to see and does not shows SD or average. You can't understand why the violin plots for needles and twigs/branches stretches so far but the box plot shows this.

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

> Answer: 1. Needles 2. Mixed 3. Twigs/Branches