Assignment 3: Data Exploration

David Robinson

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.

```
#Check working directory
getwd()
```

```
## [1] "C:/Users/dhr20/OneDrive - Duke University/1 - Academics/1 - First Year/1 - Fall 2023/2 - Enviro
```

```
#Load necessary packages
#install.packages("tidyverse")
library(tidyverse)
#install.packages("lubridate")
library(lubridate)
```

```r
#Upload the two datasets in question
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",
                    stringsAsFactors = TRUE)
Litter <- read.csv("./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: The very existence and use of insecticides (e.g., neonicotinoids) is driven by insects who are percieved as "pests" as they interact with (perhaps by eating) crops in agriculture. One reason why we might be interested in the ecotoxicology of neonicotinoids is that they could be affecting insects other than those considered pests in a given agricultural scenario. These other, non-pest insects may even be beneficial to human needs by functioning as natural predators for pests or as pollinators for crops. Source: https://www.pnas.org/doi/10.1073/pnas.2017221117

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: Quoting directly from the below source: "Woody debris is an important part of forest and stream ecosystems because it has a role in carbon budgets and nutrient cycling, is a source of energy for aquatic ecosystems, provides habitat for terrestrial and aquatic organisms, and contributes to structure and roughness, thereby influencing water flows and sediment transport (Harmon and others 1986)." Source: https://www.fs.usda.gov/research/treesearch/20001#:~:text=Woody%20debris%20is%20an%20important,influencing%20water%20flows%20and%20sediment

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 defined as material that is dropped from the forest canopy and has a butt end diameter <2cm and a length <50 cm; this material is collected in elevated PVC traps where each is a 0.5m^2 square with mesh 'basket' elevated ~80cm above the ground. Fine wood debris is defined as material that is dropped from the forest canopy and has a butt end diameter <2cm and a length >50 cm; this material is collected in ground traps as longer material is not reliably collected by the elevated traps. Ground traps are 3 m x 0.5 m rectangular areas. 2.Litter and fine woody debris sampling is executed at terrestrial NEON sites that contain woody vegetation >2m tall. Along with most of NEON's plant productivity measurements, sampling for this product occurs only in tower plots. Locations of tower plots are selected randomly within the 90% flux footprint of the primary and secondary airsheds (and additional areas in close proximity to the airshed, as necessary to accommodate sufficient spacing between plots). 3. Ground traps are sampled once per year. Target sampling frequency for elevated traps varies by vegetation present at the site, with frequent sampling (1x every 2weeks) in deciduous forest sites during senescence, and infrequent year-round sampling (1x every 1-2 months) at evergreen sites.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) #Neonics has 4623 rows and 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) #Produce summary of Effect column within Neonics data
```

```
##     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: "Mortality" and "Population" are the most common effects that are studied. These effects are of interest given that they could be indicators for the potential of neonicotinoids to harm beneficial insects and negatively impact agricultural practices.

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)) #Produce sorted summary
```

```
##                          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
```

Answer: The six most commonly studied species (other than "(Other)" which is not a species name) are Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. These species are all bees or wasps and all have a role in the pollinator ecosystem, which is of great importance for maximizing crop yield.

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) #Check the class of 'Conc.1.Author'
```

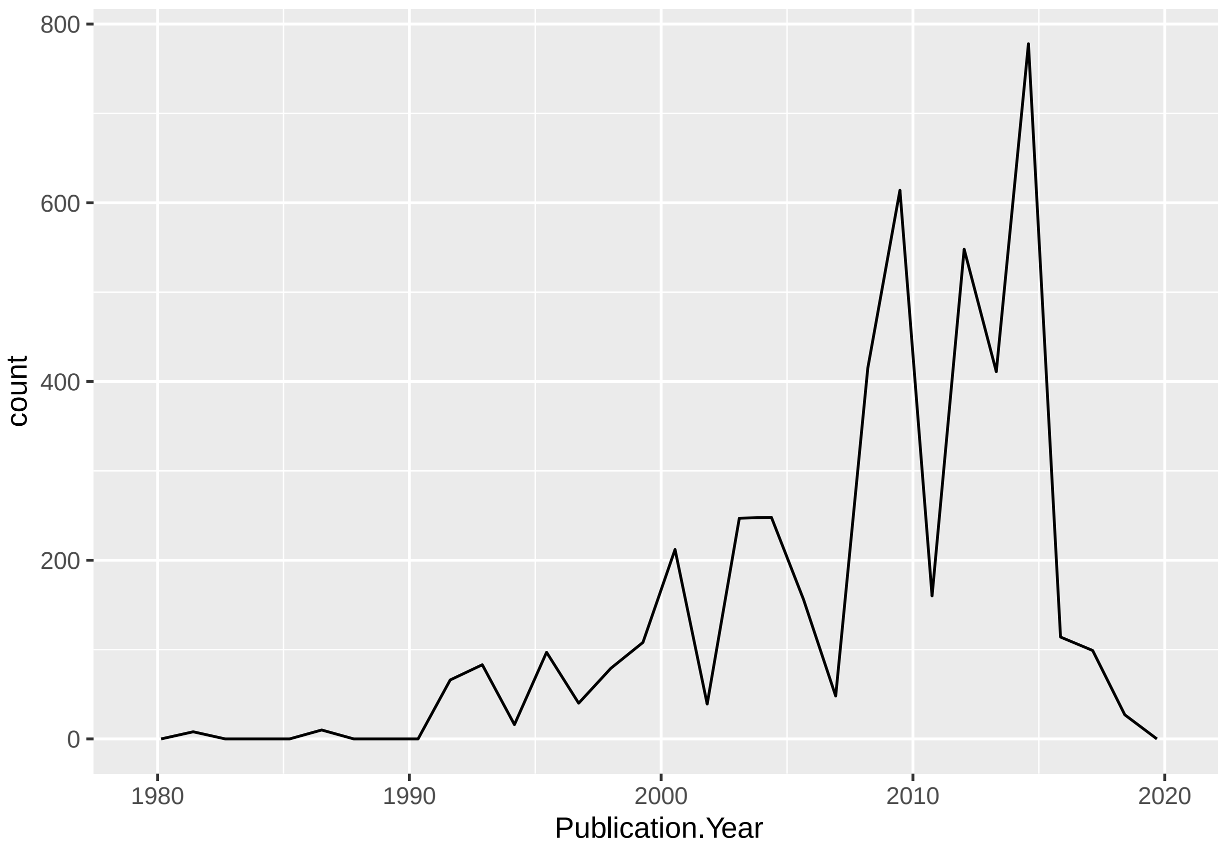
```
## [1] "factor"
```

Answer: The class is "factor." In this column of the dataset, there are characters that are non-numeric such as "NA" and numbers with "/" characters. R stores these as values to allow them to be interpreted in a manner like that of categorical variables versus solely as numbers.

Explore your data graphically (Neonics)

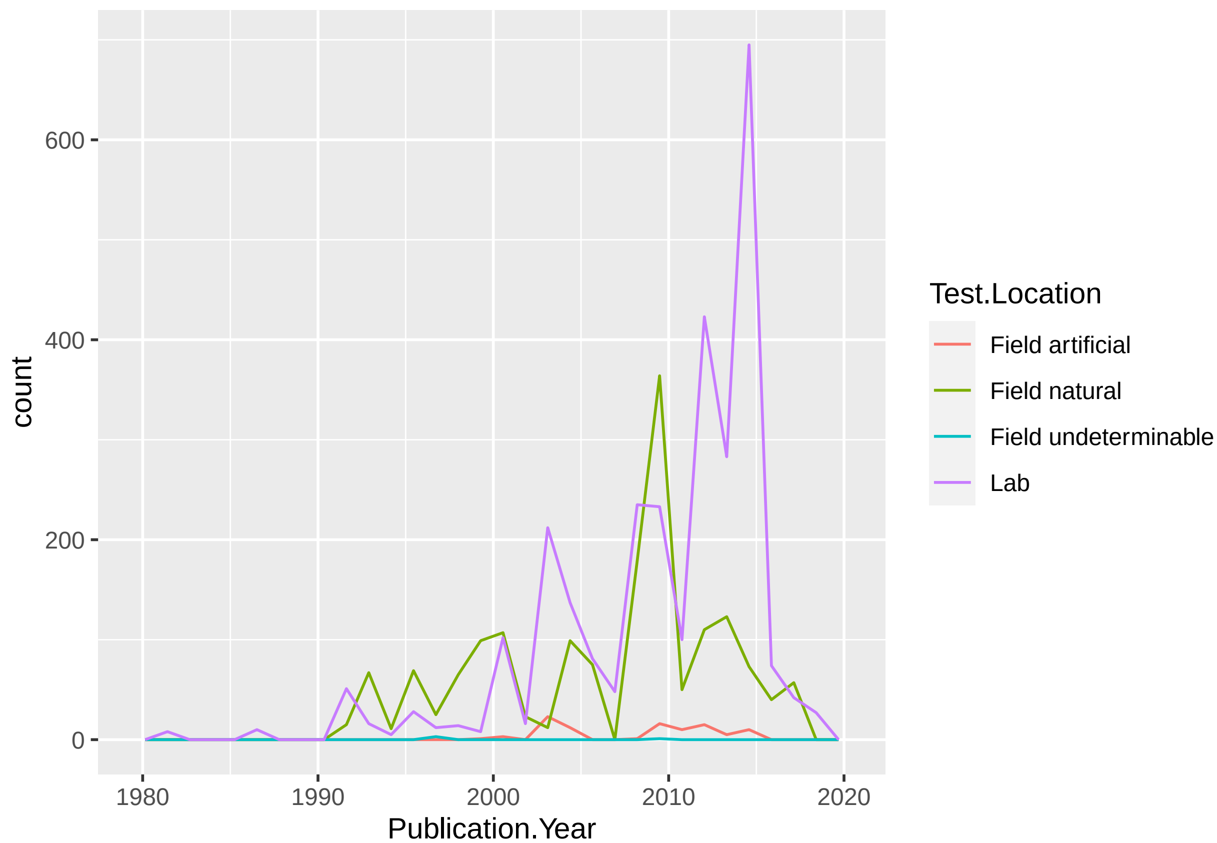
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```r
#Generate a plot of the number of studies conducted by publication year with 30 bins
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 30)
```



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

```r
#Generate the same plot with different test locations displayed as different colors
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 30)
```



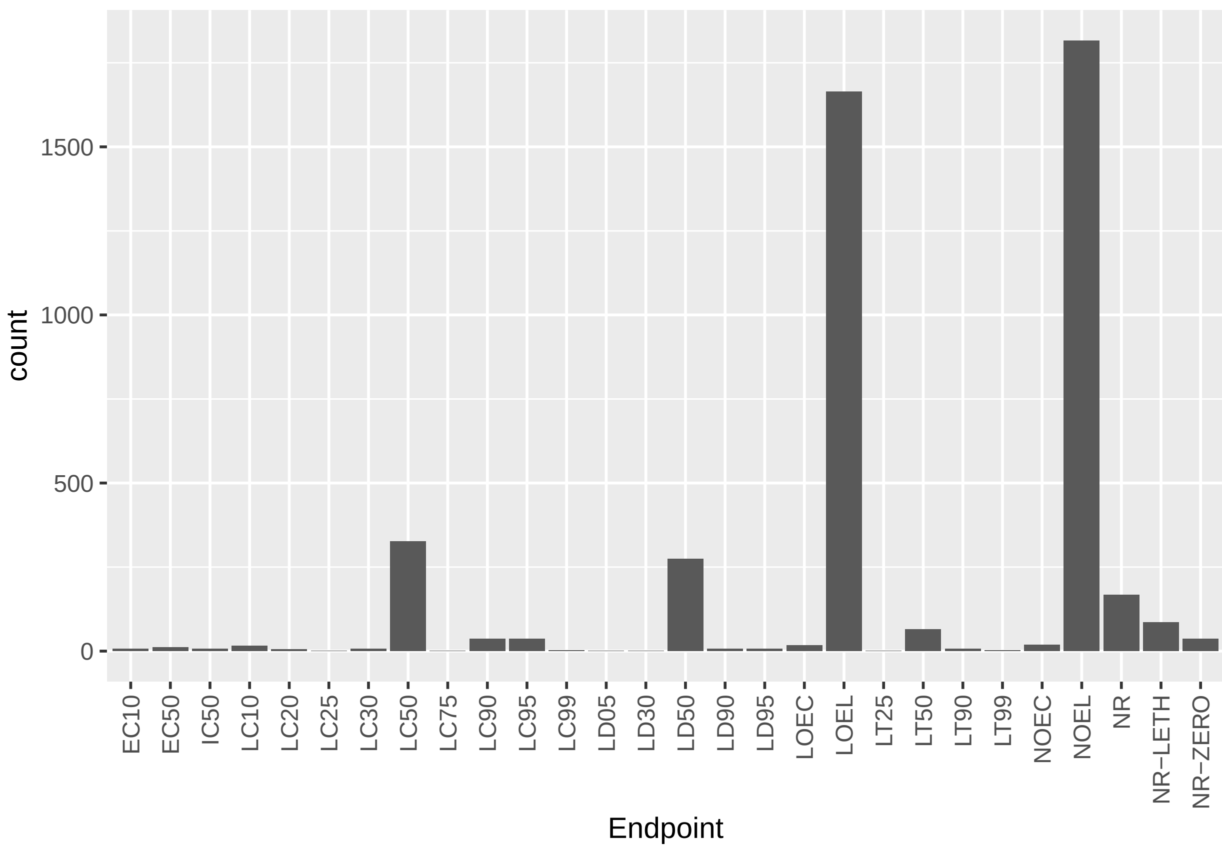
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: "Lab" and "Field natural" are the two most common test locations. While these are clearly the two most common test locations overall since 1990, there is a brief moment in ~2003 in which "Field artificial" appears to be more common than "Field natural." Lab and Field natural do differ over time – e.g., Field natural is more common than Lab from ~1993 - 2000, but then Lab is more common until ~2008, then briefly back to Field natural. After 2010, Lab is most common.

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…]

```r
#Generate a bar graph of the endpoint counts
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: The two most common points are "NOEL" and "LOEL". Per ECOTOX_CodeAppendix... "NOEL" is defined as "No-observable-effect-level: highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test." "LEOL" is defined as "Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls."

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$collectDate) #The class of collectDate was initially a factor
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
class(Litter$collectDate) #The class of collectDate is confirmed as a date
```

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

```
unique(Litter$collectDate) #Litter was sampled on the 2nd and the 30th of August 2018
```

```
## [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
unique(Litter$plotID) #Determine how many unique plots were sampled
```

```
##  [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
```

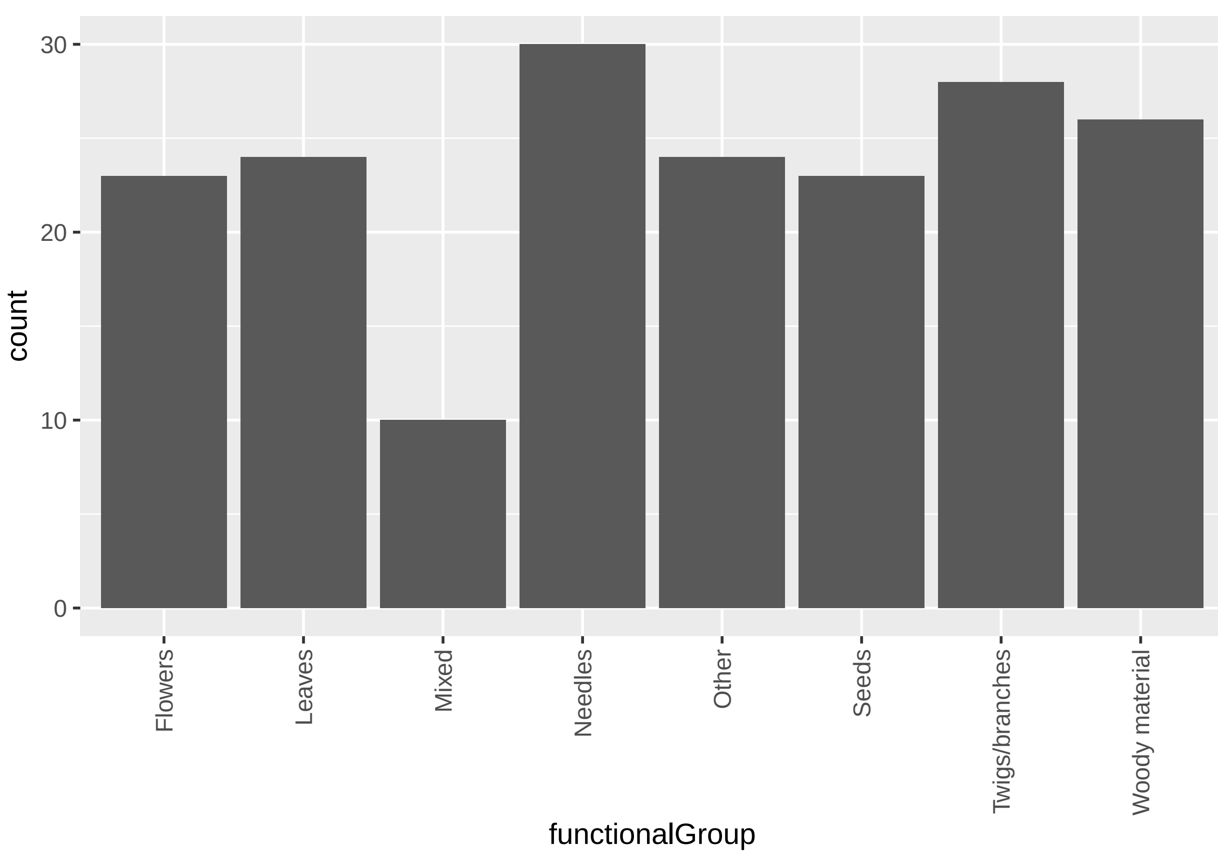
```r
summary(Litter$plotID) #Produce summary of how many times each unique plot was sampled
```

```
## 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: 12 plots were sampled at Niwot Ridge. The 'unique' function provides a list of the unique values within a set and tells us how many levels there are. The 'summary' function provides a count of the frequency of each of those unique values within a set.

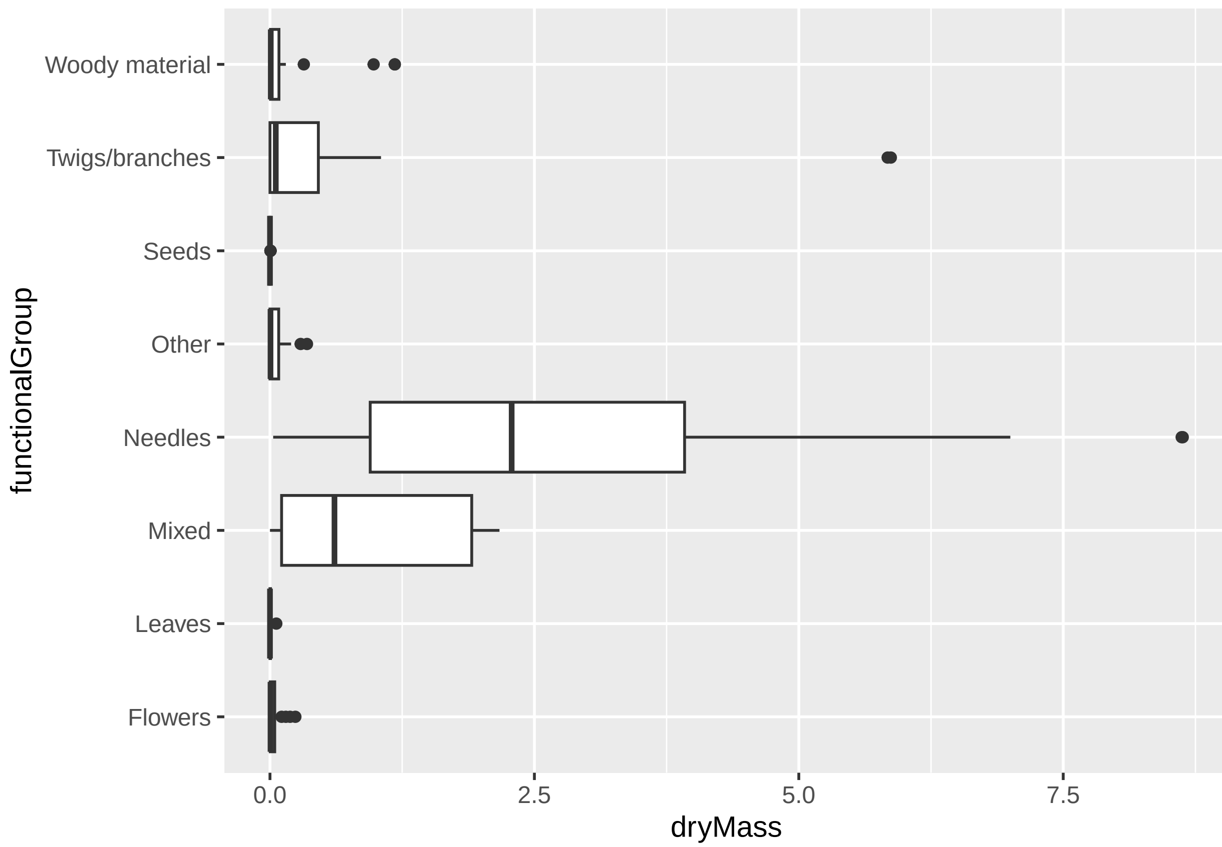
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
#Generate a bar graph of the functional group counts
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

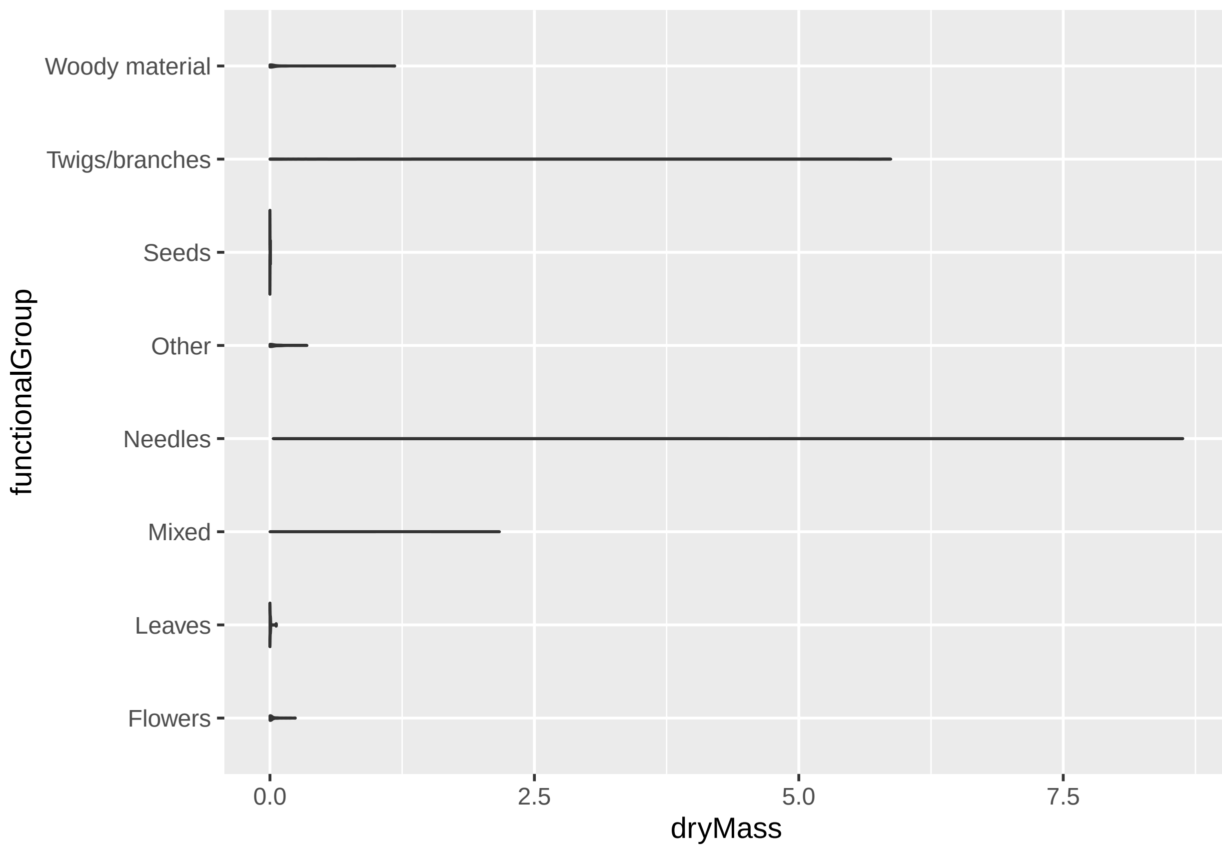


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#Generate a boxplot of dry mass by functional group
ggplot(Litter) +
  geom_boxplot(aes(x = dryMass, y = functionalGroup))
```




```
#Generate a violin plot of dry mass by functional group
ggplot(Litter) +
  geom_violin(aes(x = dryMass, y = functionalGroup))
```



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

> Answer: A box plot allows us to look at a range of values. A violion plot allows us to look at both a range of values and a distribution of values within that range. In this case, a violin plot allows us to see very little, which suggests that the distribution within the range of values is not concentrated / dense in any specific area(s). Given this feature of the data, the boxplot allows us to more clearly see the range of the data standalone including quartiles and outliers.

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

> Answer: Per the boxplot, Needles and Mixed litter tend to have the highest biomass. Their median biomass values are higher than the maximum value of most other litter types (except for Twigs/branches, which has a maximum biomass value that is greater than the median value for Mixed).