Assignment 3: Data Exploration

Caroline Rowley

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. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.
6. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai.

The completed exercise is due on Sept 30th.

Set up your R session

1. Check your working directory, load necessary packages (tidyverse), 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] "C:/Users/Caroline/Desktop/EDA-Fall2022"
```

```
setwd("C:/Users/Caroline/Desktop/EDA-Fall2022")
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library("lubridate")
```

```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
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)

# This chunk of code loads necessary packages, downloads and renames the
# desired csv files, and commands R to read the strings in as factors. I also
# had to set my working directory.
```

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: Ecotoxicology of neonicotinoids on insects is important for environmentalist and conservationist because the neonicotinoids can have a negative impact on insects life cycles. It can result in death, alter behavior and/or reduce reproduction. This can result in population decline. Additionally, insects exposed to neonicotinoids can impact the food chain and the health of the animals that consume insects (and the animals that consume those animals and so on).

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 and litter are important parts of forest ecosystems because they play a role in carbon budgets and nutrient cycling in forest ecosystems. Both can impact the fuel load and how fires might function inside of an ecosystem. They also impact the underlying soil and can alter nitrogen availability, which can influence flora composition.

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.Sampling is executed at terrestrial NEON sites that contain woody vegetation >2m tall 2. Sampling occurs only in tower plots, which are selected randomly within the 90% flux footprint of the primary and secondary airsheds 3. In forested sites, the sample plots are 20 40m x 40m plots. In sites with low-statured vegeation, sampling is targeted to 4 40m x 40m plots and 2 20m x 20m plots.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

```
## [1] 4623   30
```

```
# this code asks for the dimensions of the Neonics dataset. There are 4623
# observations and 30 variables.
```

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)
```

```
##     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
```

```
# this code shows the summary of the effects column in the Neonics dataset.
```

Answer: Population is the most common effect studied. This is probably because it gives a sense of the overall health of the population in terms of abundance and population trends. The second most common effect studied is Motality, which is related to population. Addtionally, these two effect categories are probably relatively "easy" to study.

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.

```
summary(Neonics$Species.Common.Name)
```

```
##                          Honey Bee                     Parasitic Wasp 
##                                667                                285 
##              Buff Tailed Bumblebee                Carniolan Honey Bee 
##                                183                                152 
##                         Bumble Bee                   Italian Honeybee 
##                                140                                113 
##                    Japanese Beetle                  Asian Lady Beetle 
##                                 94                                 76 
##                     Euonymus Scale                           Wireworm 
##                                 75                                 69
##                  European Dark Bee                  Minute Pirate Bug
##                                 66                                 62
##               Asian Citrus Psyllid                      Parastic Wasp
##                                 60                                 58
##             Colorado Potato Beetle                    Parasitoid Wasp
##                                 57                                 51
##                Erythrina Gall Wasp                       Beetle Order
##                                 49                                 47
##        Snout Beetle Family, Weevil           Sevenspotted Lady Beetle
##                                 47                                 46
##                     True Bug Order              Buff-tailed Bumblebee
##                                 45                                 39
##                       Aphid Family                     Cabbage Looper
##                                 38                                 38
##               Sweetpotato Whitefly                      Braconid Wasp
##                                 37                                 33
##                       Cotton Aphid                     Predatory Mite
##                                 33                                 33
##             Ladybird Beetle Family                         Parasitoid
##                                 30                                 30
##                      Scarab Beetle                      Spring Tiphia
##                                 29                                 29
##                        Thrip Order               Ground Beetle Family
##                                 29                                 27
##                 Rove Beetle Family                      Tobacco Aphid
##                                 27                                 27
##                       Chalcid Wasp             Convergent Lady Beetle
##                                 25                                 25
##                      Stingless Bee                  Spider/Mite Class
##                                 25                                 24
##                Tobacco Flea Beetle                   Citrus Leafminer
##                                 24                                 23
##                    Ladybird Beetle                          Mason Bee
##                                 23                                 22
##                           Mosquito                      Argentine Ant
##                                 22                                 21
##                             Beetle         Flatheaded Appletree Borer
##                                 21                                 20
##               Horned Oak Gall Wasp                 Leaf Beetle Family
##                                 20                                 20
##                  Potato Leafhopper         Tooth-necked Fungus Beetle
##                                 20                                 20
##                       Codling Moth          Black-spotted Lady Beetle
##                                 19                                 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
##              Araneoid Spider Order                          Bee Order
##                                 17                                 17
##                      Egg Parasitoid                       Insect Class
##                                  17                                 17
##            Moth And Butterfly Order       Oystershell Scale Parasitoid
##                                  17                                 17
## Hemlock Woolly Adelgid Lady Beetle              Hemlock Wooly Adelgid
##                                  16                                 16
##                                Mite                        Onion Thrip
##                                  16                                 16
##               Western Flower Thrips                       Corn Earworm
##                                  15                                 14
##                   Green Peach Aphid                          House Fly
##                                  14                                 14
##                           Ox Beetle                 Red Scale Parasite
##                                  14                                 14
##                  Spined Soldier Bug              Armoured Scale Family
##                                  14                                 13
##                    Diamondback Moth                      Eulophid Wasp
##                                  13                                 13
##                   Monarch Butterfly                      Predatory Bug
##                                  13                                 13
##               Yellow Fever Mosquito                Braconid Parasitoid
##                                  13                                 12
##                        Common Thrip       Eastern Subterranean Termite
##                                  12                                 12
##                              Jassid                         Mite Order
##                                  12                                 12
##                           Pea Aphid                   Pond Wolf Spider
##                                  12                                 12
##            Spotless Ladybird Beetle             Glasshouse Potato Wasp
##                                  11                                 10
##                            Lacewing            Southern House Mosquito
##                                  10                                 10
##             Two Spotted Lady Beetle                         Ant Family
##                                  10                                  9
##                        Apple Maggot                            (Other)
##                                   9                                670
```

```
# this chunk of code shows the summary of species in the Neonics dataset.
```

Answer: The six most common species are the Honey Bee, the Parasitic Wasp, the Buff Tailed Bumblebee, the Carniolan Honey Bee, the Bumble Bee, and the Italian Honeybee. Many of these species are closely related (bees), and the Parasitic Wasp is considered to be a "beneficial" insect similar to many bee species. They are likely of interest to this dataset because of the important role they play in the ecosystem as pollinators.

8. Concentrations are always a numeric value. What is the class of Conc.1..Author. in the dataset, and why is it not numeric?

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

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

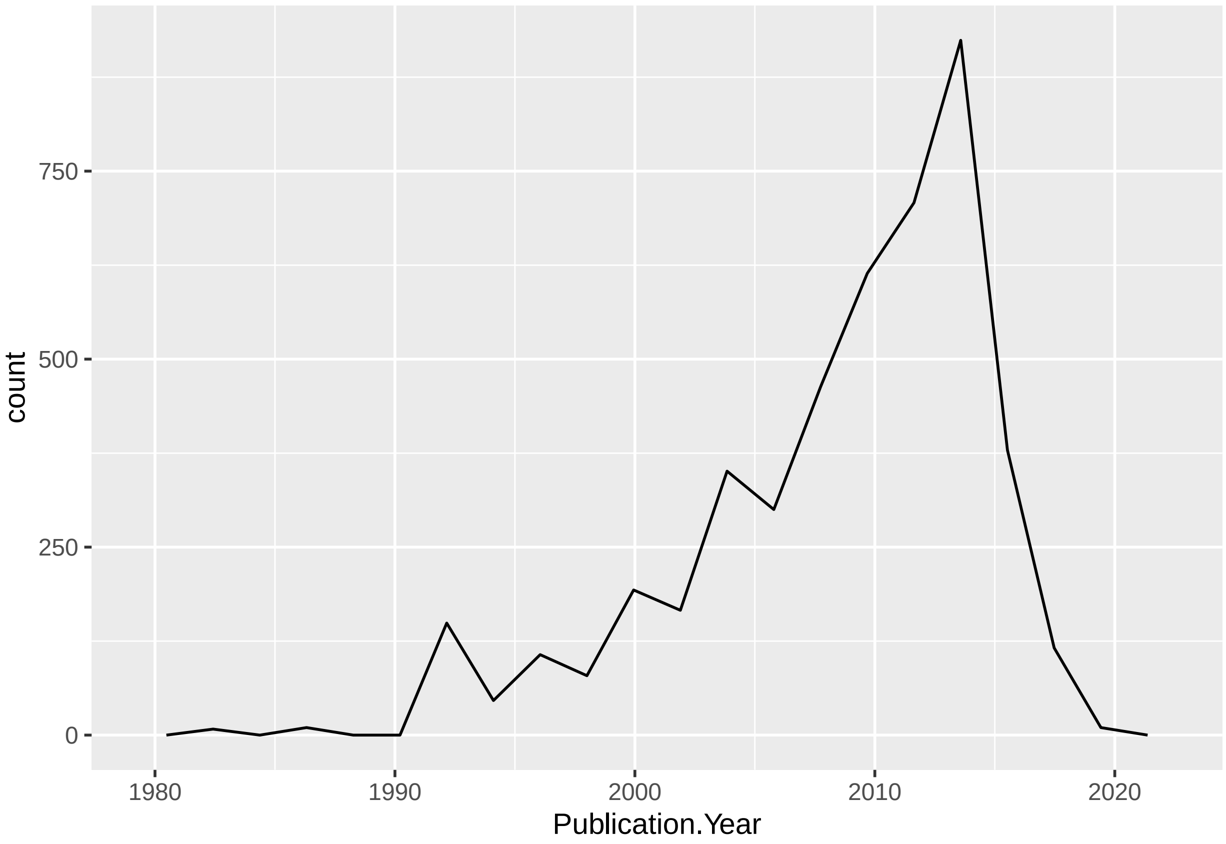
```
# this code is asking for the class associated with the Conc.1..Author column
# in the Neonics dataset.
```

> Answer: The class is a factor, likely because the column has approximations, less than and greater than symbols in the column's values. The computer reads these symbols and classifies the column as a factor because it has both numeric and non-numeric values.

Explore your data graphically (Neonics)

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

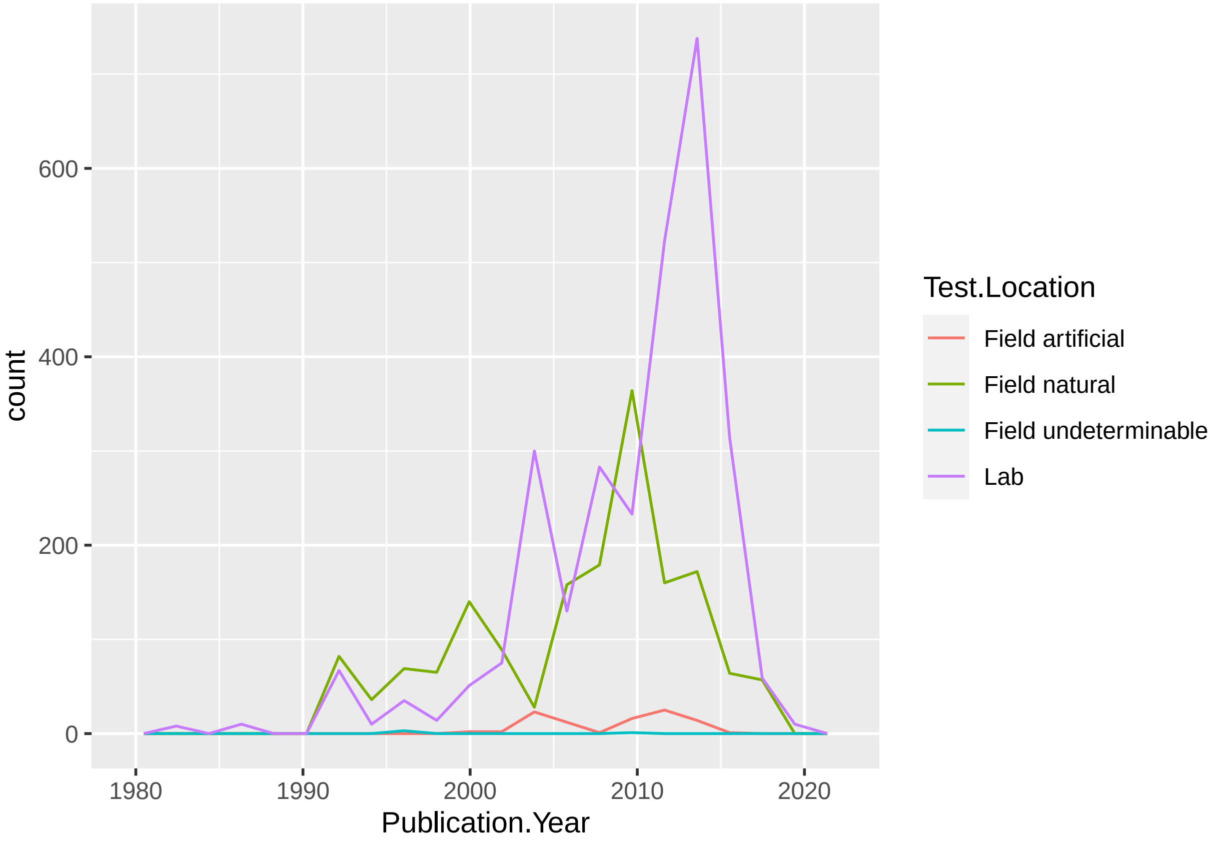
```
ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year), bins = 20)
```



```
# This code asks ggplot to chose the Neonics dataset, and then create a
# frequency polygon of the publication year. It also sets the bins to 20.
```

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),
    bins = 20)
```



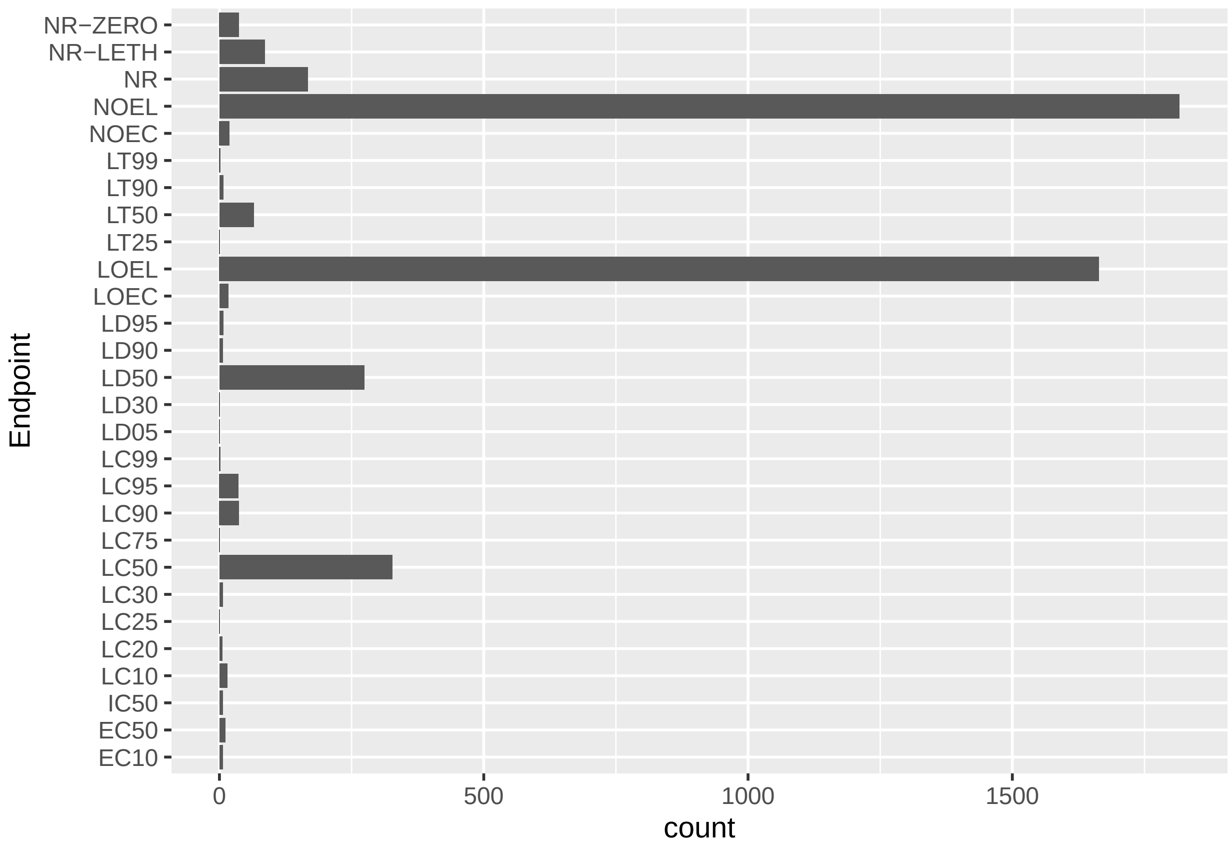
```
# This code asks ggplot to chose the Neonics dataset, create a frequency
# polygon of the publication years, and create different plots for each Test
# Location. It also sets the bins to 20.
```

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

> Answer: The most common test location appears to the be lab, with a significant number of tests being run in 2014. Field natural appears to be the 2nd most common test location. Field Artifical is far less common, and Field undeterminable is the least 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.

```
ggplot(Neonics) + geom_bar(aes(x = Endpoint)) + coord_flip()
```



```
# This code asked ggplot to create a bar graph from the Neonics dataset of the
# endpoints I flipped the axis so that the end point categories can be read
```

> Answer: The two most common end points are NOEL (No observable effect level: highest dose (concentration)- the test produced effects that were not significantly different from responses of controls) and LOEL (Lowest observable effect level: lowest dose(concentration)- the test produced effects that were significantly different 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)
```

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

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

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

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

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

```
# This code checks the class of the collectDate column in the Litter dataset.
# It then informs R that this column is a date, and the format of the date
# inside that column. It then asks R what the unique values are in the
# collectDate column.
```

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`?

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

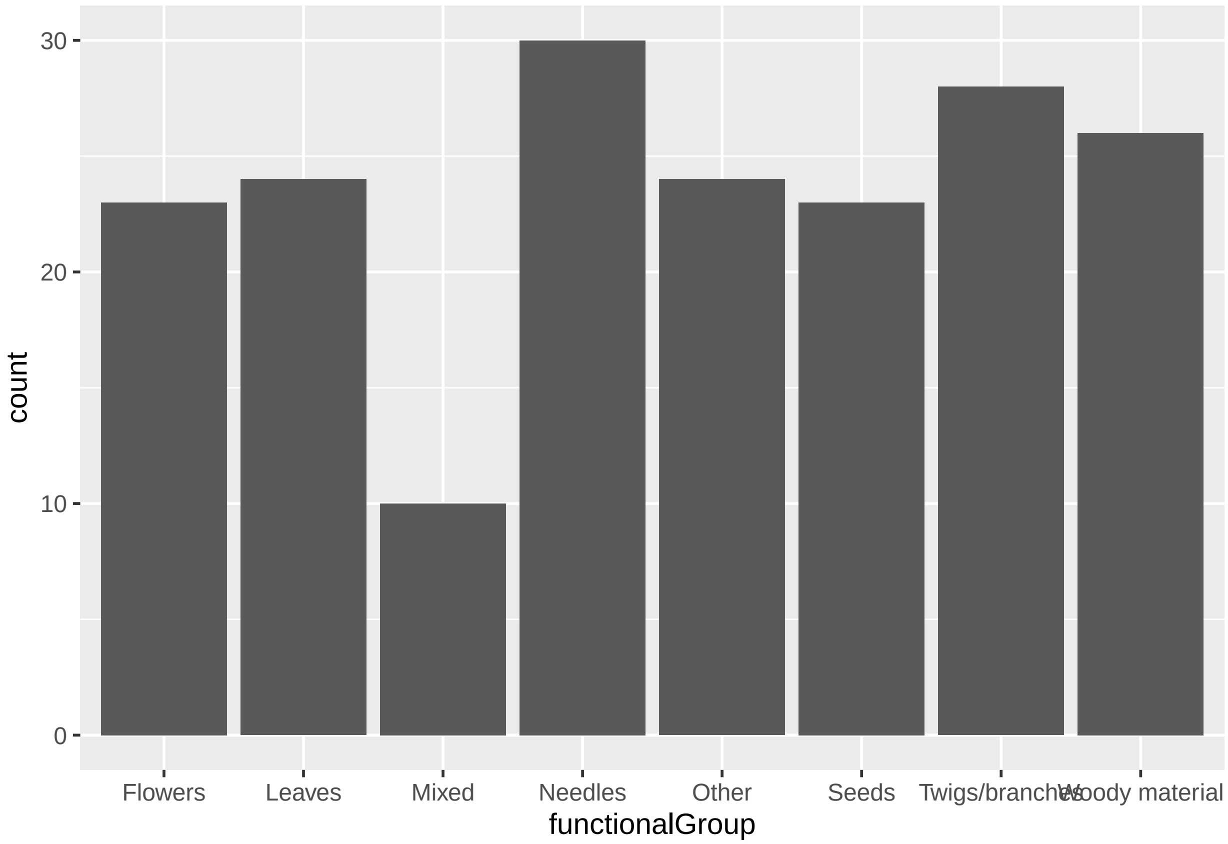
```
##  [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

```
# This code asks for the unique values in the plotID column of the Litter
# dataset
```

> Answer: There are 12 plots sampled in this data set. The information obtained from unique is different than summary because the unique command will tell you how many different values are present (in this case, 12), whereas summary would have told how many samples were taken at each plot.

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.

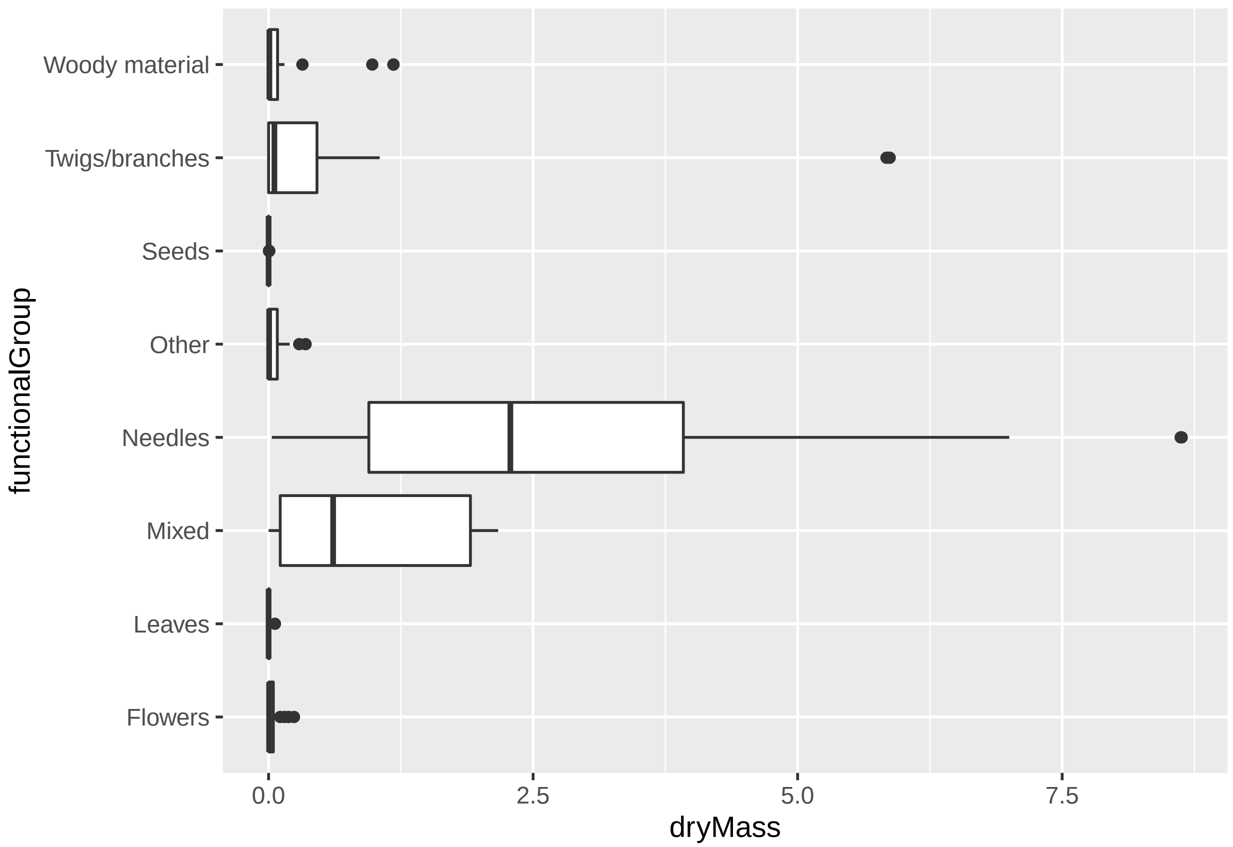
```
ggplot(Litter) + geom_bar(aes(x = functionalGroup))
```



```
# this code asks ggplot to make a bar plot of the functionalGroup column from
# the Litter dataset.
```

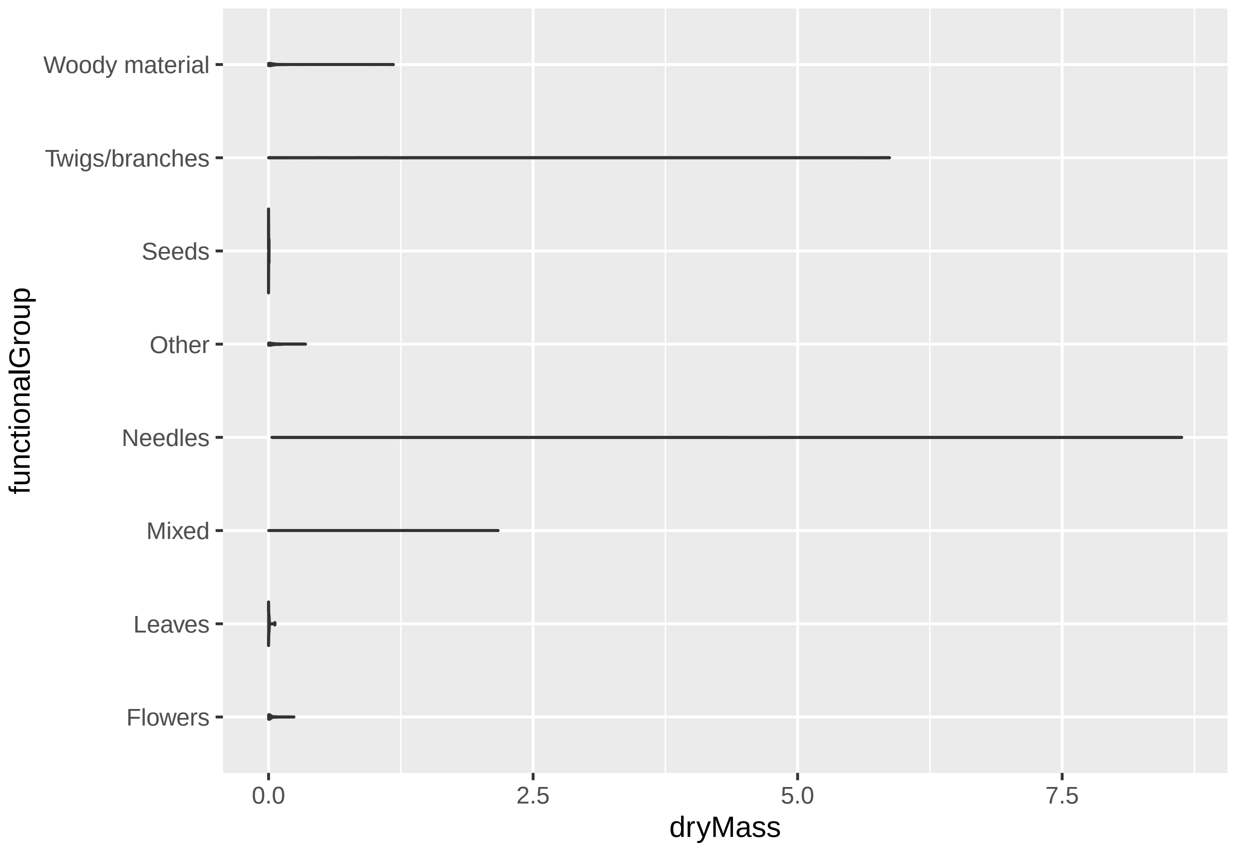
15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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



```
# this code asks ggplot to make a boxplots of the dryMass column and to sort
# the boxplots by functionalGroup

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



```
# this code asks ggplot to make a violin plot of the dryMass column and to sort
# the violin plots by functionalGroup
```

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

> Answer: Violin plots are used to show the distributions of numeric data using density curves. The width corresponds to the approximate frequency of that points in each region. In this case, the violin plots are extremely thin; our data is not robust enough to create meaninful density curves. If we had a larger data set, we might be able to extrapolate more from violin plots. Box Plots show the minimum and maximum values (the length of the line), outliers (as dots), the median (as a line inside the box) and the IQR (the length of the box). For out data set, this is a much more helpful visualization.

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

> Answer: Needles tend to have the highest biomass at these sites