Assignment 3: Data Exploration

Michael Gaffney, Section #2

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Exploration.

Directions

1. Change "Student Name, Section #" on line 3 (above) with your name and section number.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "FirstLast_A03_DataExploration.Rmd") prior to submission.

The completed exercise is due on <>.

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 add the `stringsAsFactors = TRUE` parameter to the function when reading in the CSV files.**

```
getwd()
```

```
## [1] "/Users/michaelgaffney/Documents/Duke University/Nicholas School of the Environment/05 Spring 20
```

```
library(tidyverse)
#read in the data
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 ecotoxicologoy of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: It would be important to establish how toxic these chemicals are to particular insects, to understand both how effective they might be in agricultural settings and how dangerous they might be outside of the area of application. We might want to know how well these substances kill particular pests, and whether they are dangerous to insects that we would prefer not to harm.

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: Measuring litter and woody debris would give us a good indication of how much biomass is cycling in the system. This might be important for evaluating the overall health of the ecosystem, and it could also provide a useful metric for understanding various kinds of ecosystem services like carbon sequestation.

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 occurs within what they call tower plots. There are about 20 of these per site, each 40x40m or a mix of 40x40m and 20x20m. Within each of these plots, there are roughly 1 to 4 litter trap pairs placed, one on the ground and one elevated.* 2. Ground traps (for woody debris) are sampled once per year. Elevated traps (for litter) are sampled differently. They are sampled more frequently for deciduous sites in autumn, and less frequently for evergreen sites overall. Deciduous sites sometimes are not checked during dormant season. The temporal resolution for these types of debris will therefore be different. *3. The finest temporal resolution for any sample is located in the "daysOfTrapping" column. This columnn records the range between the set date for the trap and collection date of the material; in other words, it tells us precisely how many days passed during the accumulation of the debris.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
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?

```
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: Mortality and Population are the comst commonly studied effects. This makes sense. Insecticides are designed to have population-level effects on the ability of insects to live; logically, we would be most concerned with quantifying this effect. We might be especially concerned if there are species we want to avoid harming with the insecticides, like bees.

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
```

Answer: Various forms of honeybees are the most commonly studied insect. Honeybees are extremely important in agriculture for pollination. The study is likely concerned with how dangerous this insecticide is for honeybees. Too much mortality would likely be a serious impediment to the use of this particular chemical.

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.)
```

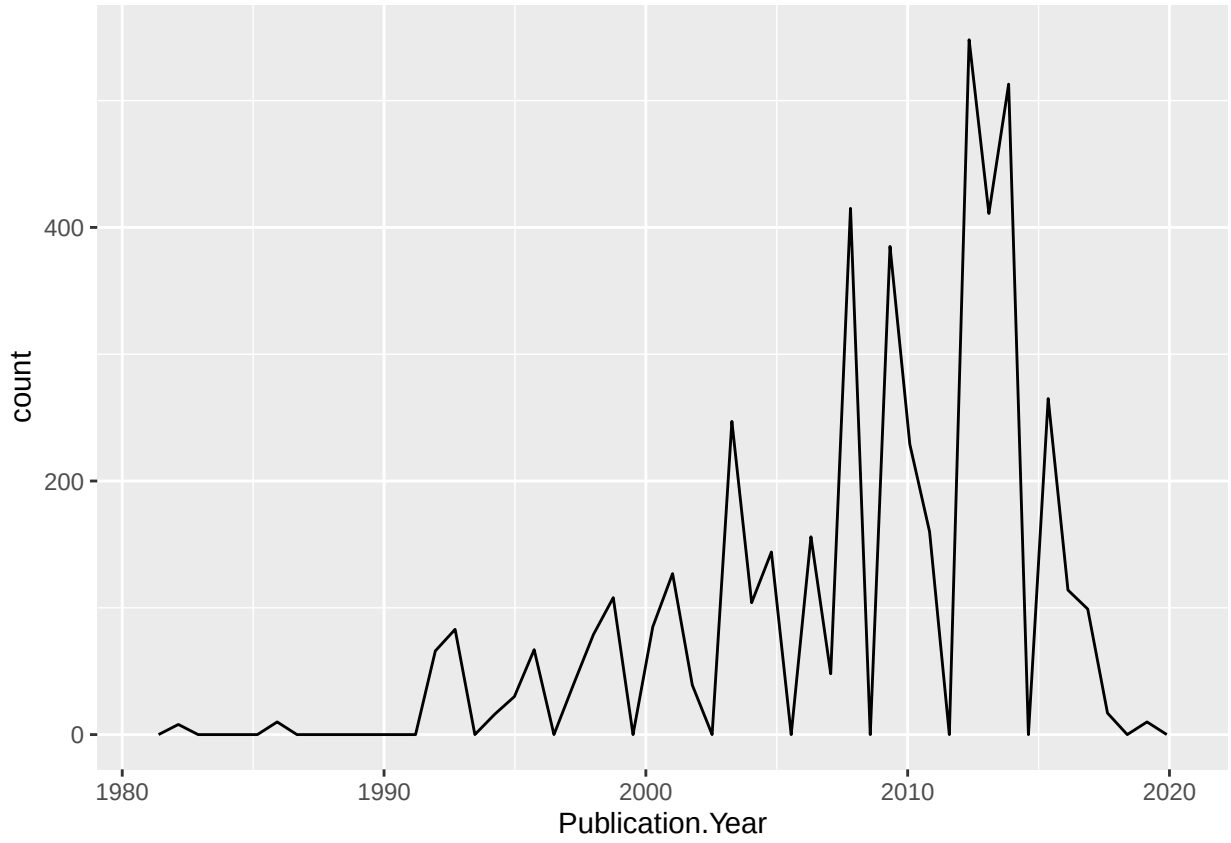
```
## [1] "factor"
```

Answer: This is a factor. R likely read this is as a factor becuase some of the values appear to not be numeric, or to have slashes next to the numbers; it would normally default to a string, but when reading in the csv I told R to treat it as a factor.

Explore your data graphically (Neonics)

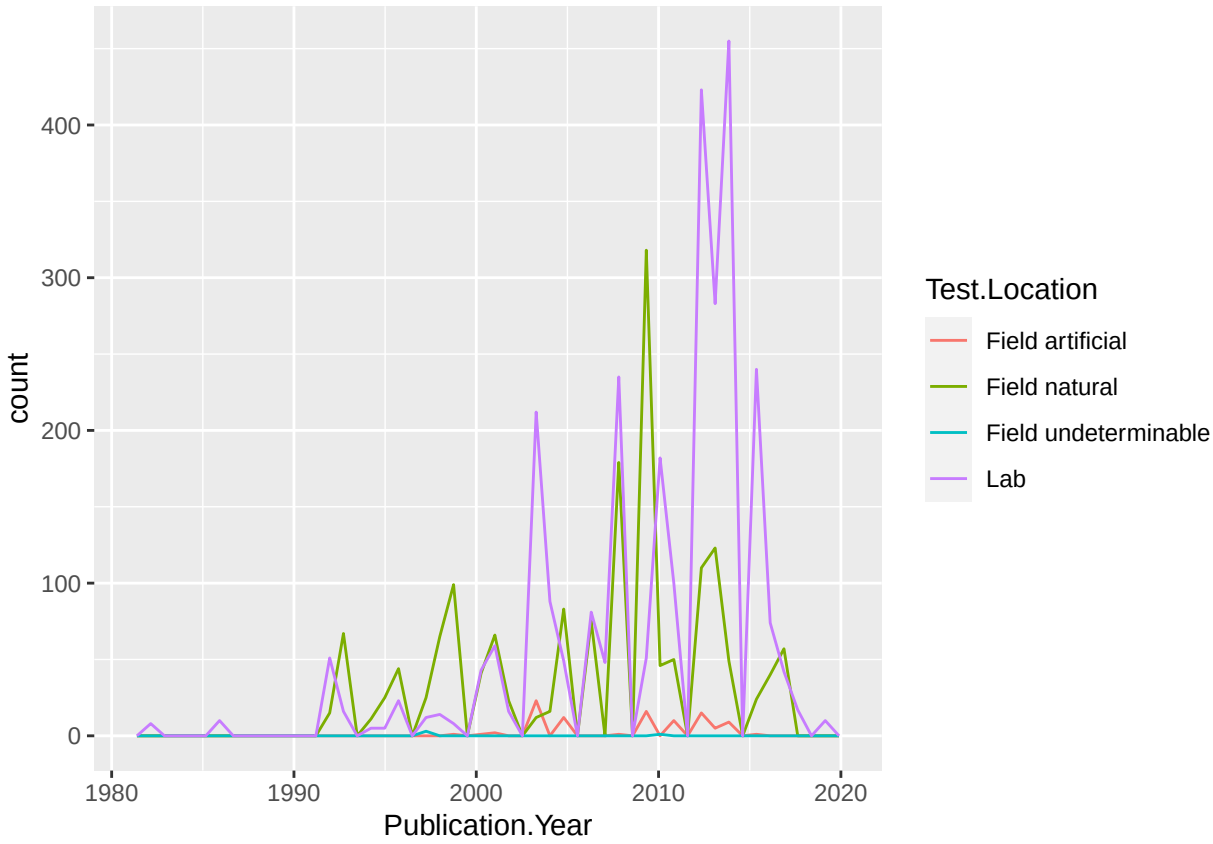
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50)
```



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 = 50)
```

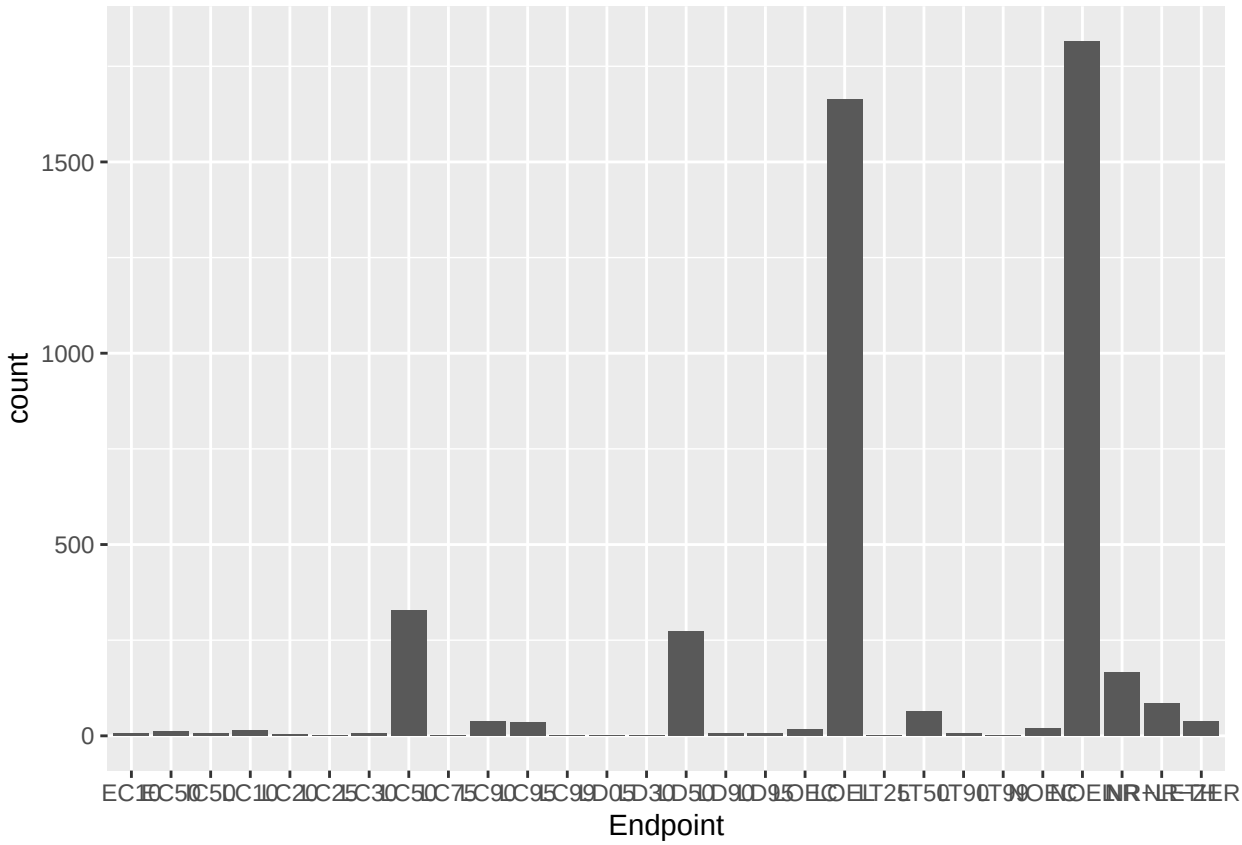


Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: Lab and natural field locations seem to be the most prevalent. These differ over time. Lab tests seem to increase steadily until about 2015, and then begin to decrease. Field tests increase until around 2010, and then begin to decline in frequency.

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(Endpoint))
```



Answer: LOEL and NOEL are the two most common. NOEL means "no observable effect level," meaning that the highest dose produced effects that were not significantly different from the control group. LOEL meanwhile refers to "lowest observable effect level," which refers to lowest dose concetrations producing an effect that was significantly different from the control response.

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.

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
class(Litter$collectDate)
```

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

```
#use lubridate to convert to date
Litter$collectDate <- ymd(Litter$collectDate)
#check class again
class(Litter$collectDate)
```

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

```
#check which dates the litter was sampled in August 2018
unique(Litter$collectDate)
```

```
## [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`?

```
unique(Litter$namedLocation)
```

```
##  [1] NIWO_061.basePlot.ltr NIWO_064.basePlot.ltr NIWO_067.basePlot.ltr
##  [4] NIWO_040.basePlot.ltr NIWO_041.basePlot.ltr NIWO_063.basePlot.ltr
##  [7] NIWO_047.basePlot.ltr NIWO_051.basePlot.ltr NIWO_058.basePlot.ltr
## [10] NIWO_046.basePlot.ltr NIWO_062.basePlot.ltr NIWO_057.basePlot.ltr
## 12 Levels: NIWO_040.basePlot.ltr ... NIWO_067.basePlot.ltr
```

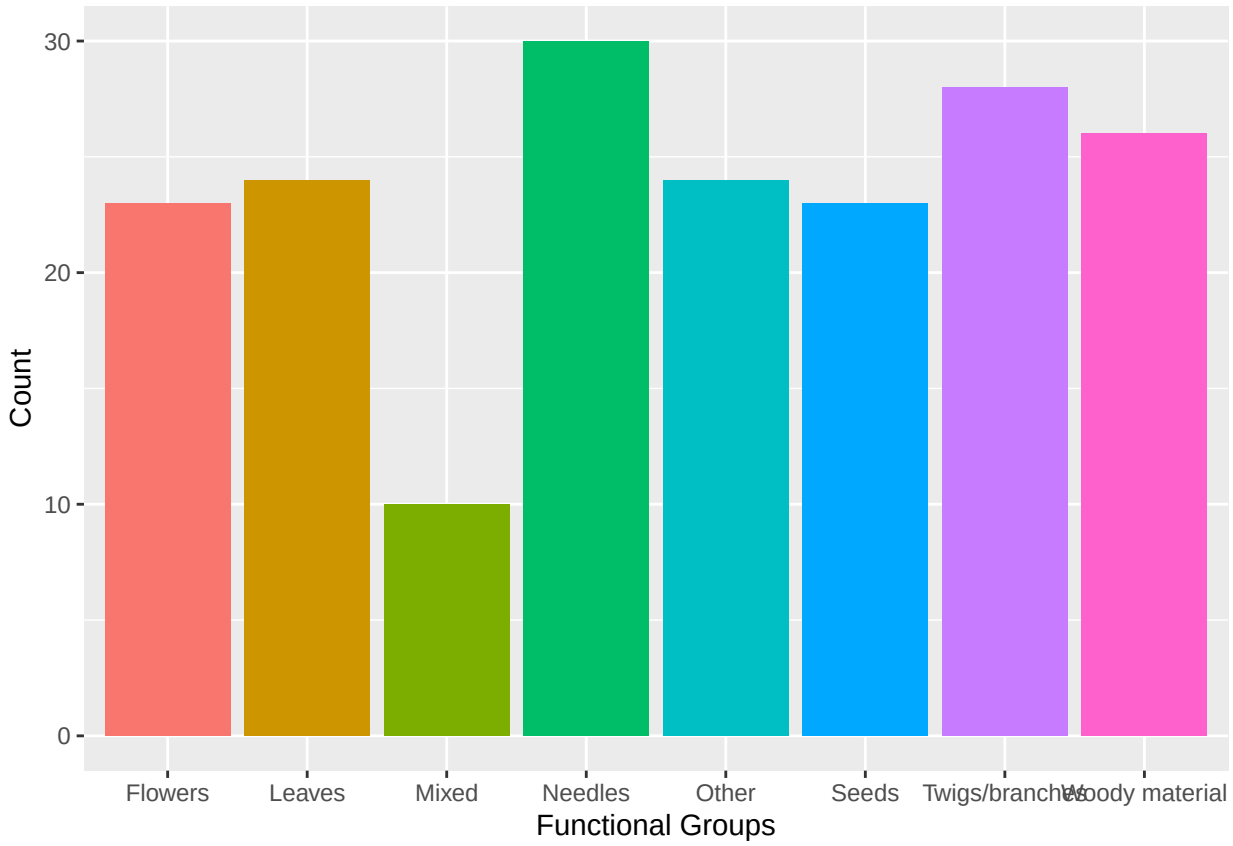
```
summary(Litter$namedLocation)
```

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

Answer: There are 12 unique sites. The unique function provides a list of the different "levels" in a column; the summary function provides a count of how many occurences there are of each value. Unique, in other words, tells you how many different values in a column, summary tells you how often those values occur.

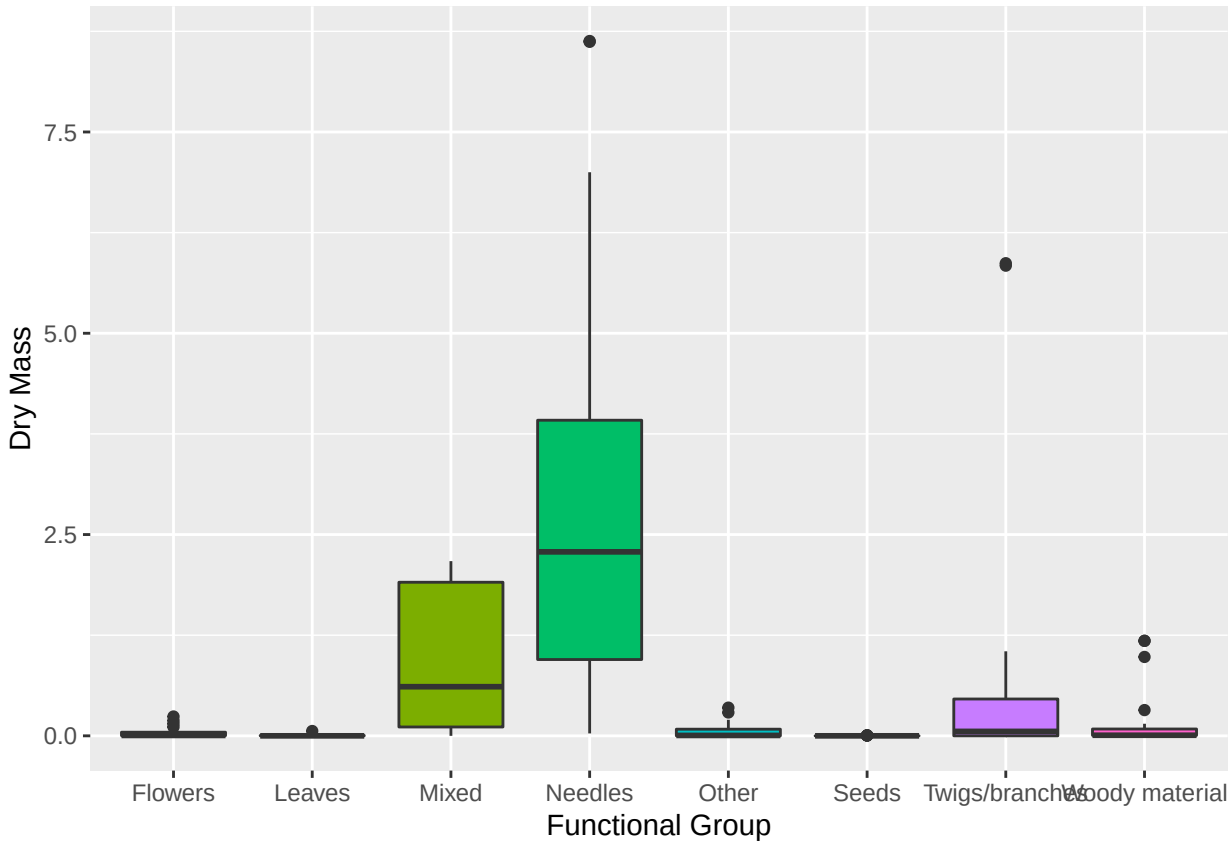
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter) +
  #add some color to the bars and remove the legend
  geom_bar(aes(functionalGroup, fill = functionalGroup), show.legend = FALSE) +
  xlab("Functional Groups") +
  ylab("Count")
```

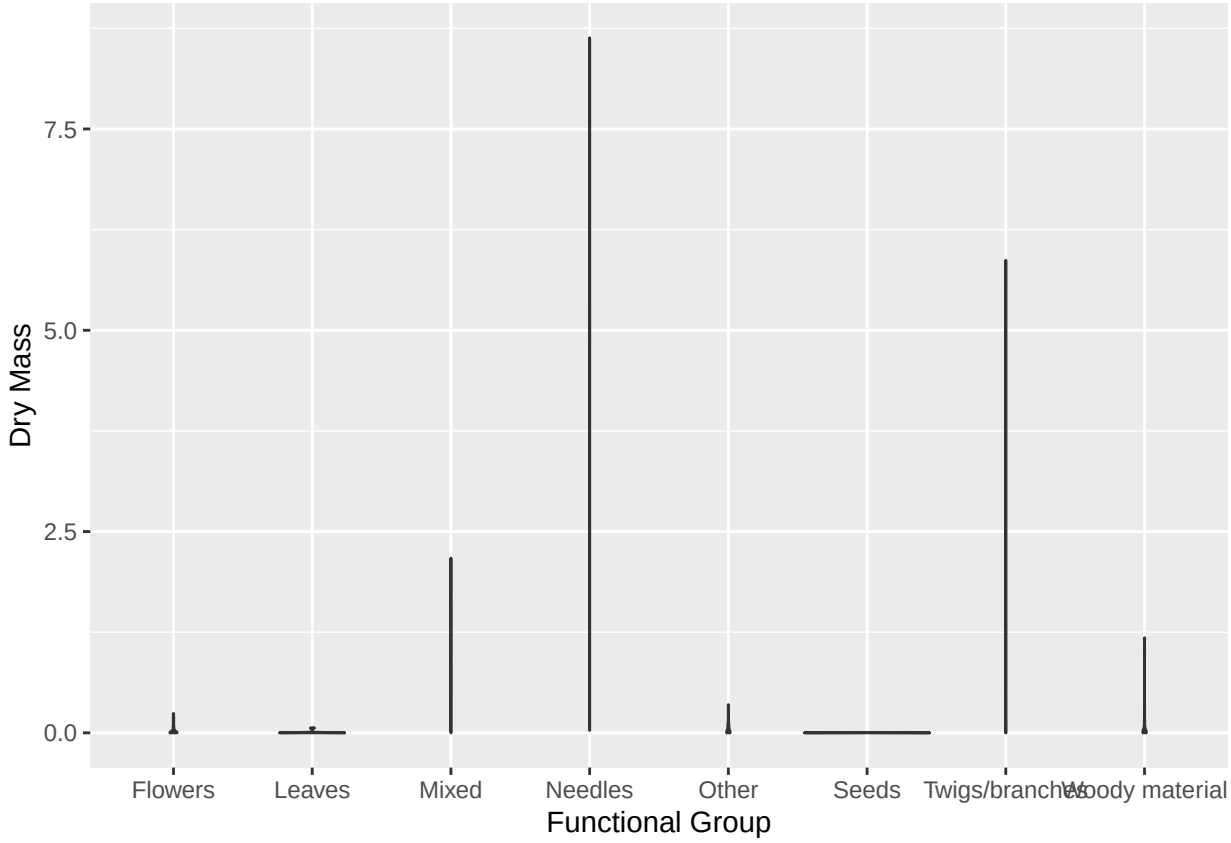


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#create boxplot of dry mass for each functional group
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass, fill = functionalGroup), show.legend = F) +
  xlab("Functional Group") +
  ylab("Dry Mass")
```




```
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass)) +
    xlab("Functional Group") + ylab("Dry Mass")
```



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

> Answer: Because of the way the data is distributed, the violin plot seems to be hiding the bulk of the data, meaning that it does not really show very effectively where mass is distributed in each category. The boxplot, by contrast, is more focused on the inter-quartile range, and therefore is more effective for this data.

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

> Answer: Needles and mixed tend to have the highest biomass. This would lead us to believe that the site is mostly pine, I would assume.