Assignment 3: Data Exploration

Jacob Freedman

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.

```
#Loading packages and reading in the datasets with relative path
library(tidyverse)
library(dplyr)
library(lubridate)

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: If we are using neonicotinoids as insectisides, it would be valuable to know their effects on insects. We would want to know the effectiveness of certain compounds, the time required to cause mortality, or other characteristics of a given chemical. We also would want to know whether a chemical would impact non-target species, such as bees, which could act as valuable pollinators.

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: Fun fact, I worked for NEON at Niwot Ridge for two seasons in 2019 and 2020 and collected some of this data (though I was really on the animal team and did more work with beetles). Quantifying and categorzing litter is important for many reasons. The type of litter can impact the acidity of the soil (pine needles) and change the area's vulnerability to fire. Different amounts/forms of litter might also impact the insect and small mammal communities that live in the area.

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. There are both elevated and ground level litter traps. 2. Litter is sorted into functional groups before being weighed. 3. The weight is a dry mass, not a wet field mass.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

```
dim(Litter)
```

```
## [1] 188  19
```

```
#Neonics is 4623 rows and 30 columns. Litter is 188 rows and 19 columns.
```

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?

```
#Using summary to determine common effects

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: The most common effects are Population (1803) and Mortality (1493), which comprise the majority of the total observations. Population is of interest because companies may be interested in understanding how the chemical impacts the numbers of a large population of insects, even if it doesn't result in direct or immediate mortality. In turn, they may be interested in mortality to assess the direct impact of the chemical on an individual organism.

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.

```
#Using summary to list most common species and then looking at first six.

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: The 6 most commonly studied species are the Honey Bee (667), Parasitic Wasp (285), Buff Tailed Bumblebee (183), Carniolan Honey Bee (152), Bumble Bee (140), and the Italian Honeybee (113). These are likely the most studied because they are non-target organisms that

companies want the pesticides to NOT adversely impact. All of the bee species are important pollinators and parasitic wasps are crucial for eliminating certain pests.

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

```
#determining the class of Conc.1..Author

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

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

Answer: It is a factor because the data stored in the columns are not all numbers. They have additional symbols that make it impossible for r to determine the class as numeric. For example, the concentration values are sometimes expressed as ">###" (greater than a given value) or are written as "###/". This may mean that only a certain number of significant figures were used and the value is approximate, but I could not find more information in the Code Appendix.

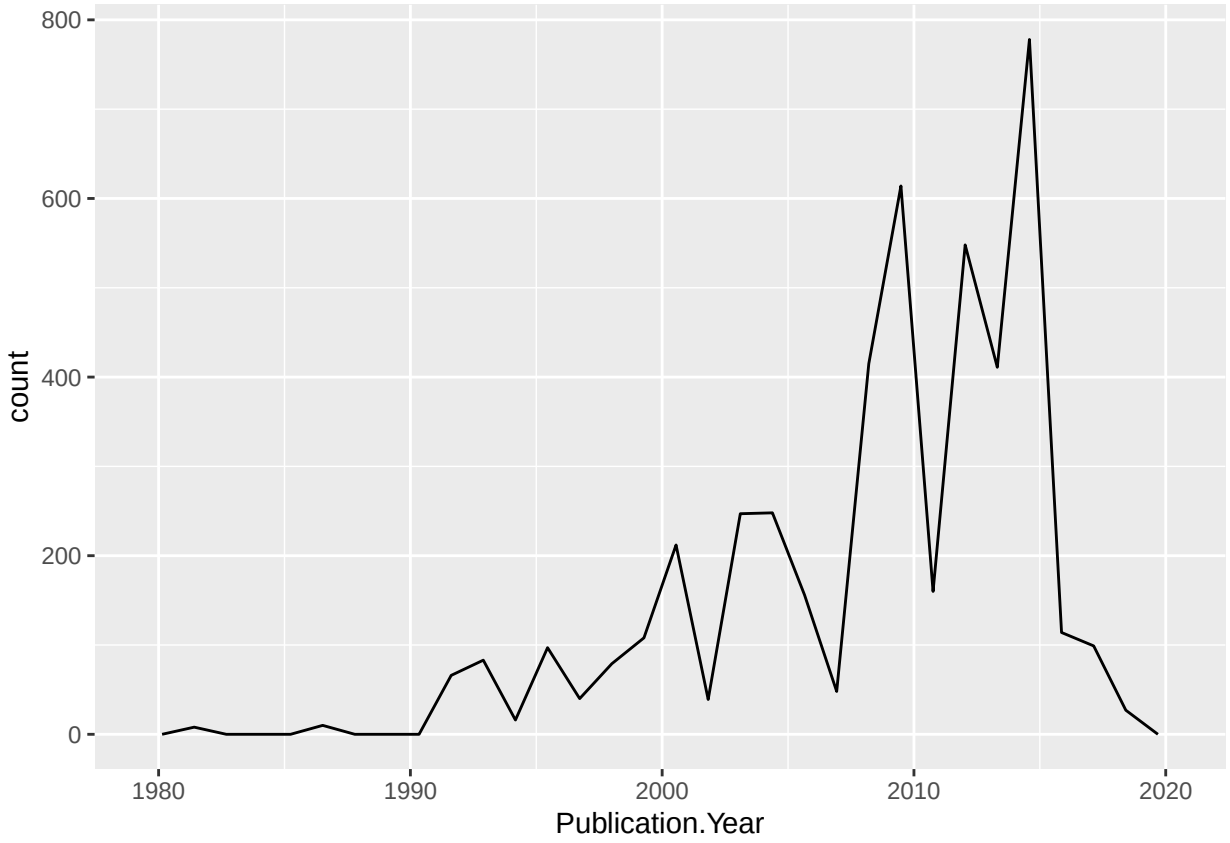
Explore your data graphically (Neonics)

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

```
#Creating a plot of Publication Year

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

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

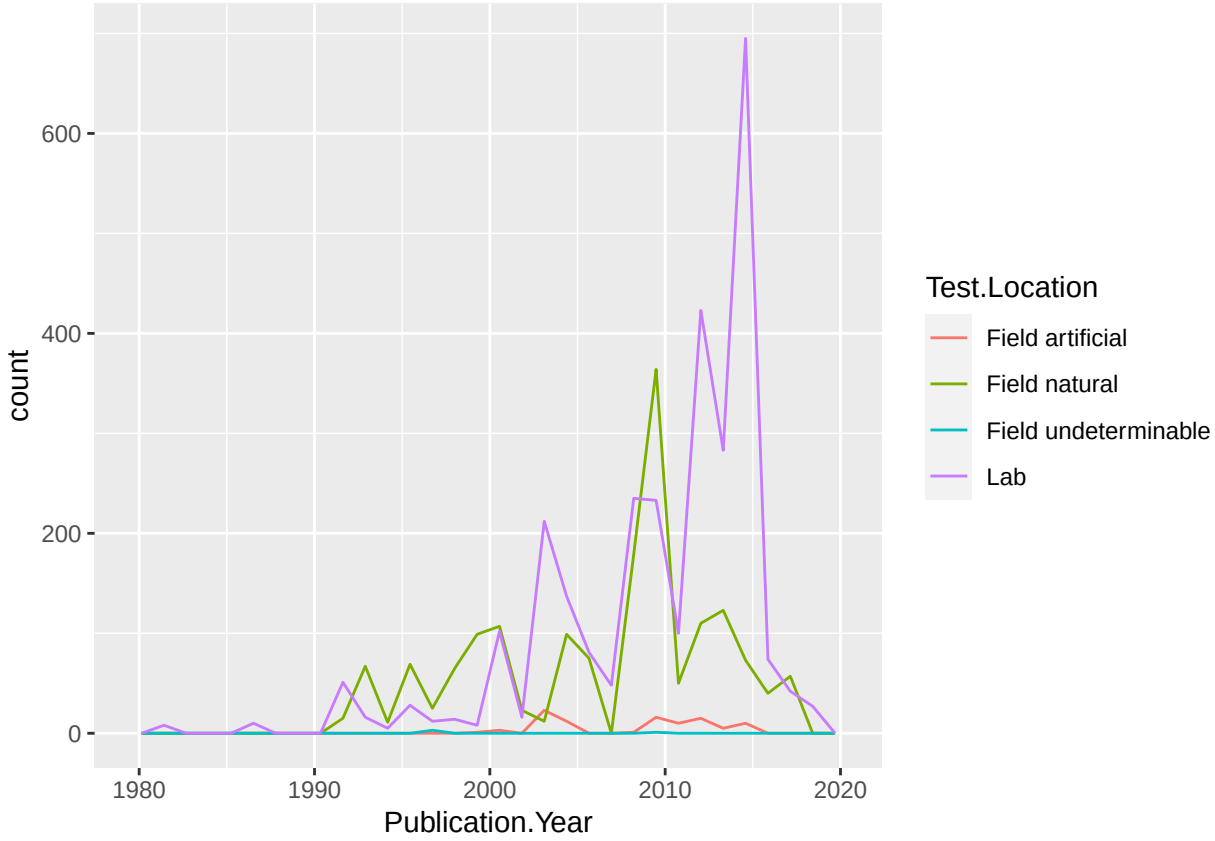


10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
#Using colors to distinguish Publication Year counts by Test Location

ggplot(Neonics, aes(Publication.Year, after_stat(count), colour = Test.Location)) +
  geom_freqpoly()
```

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



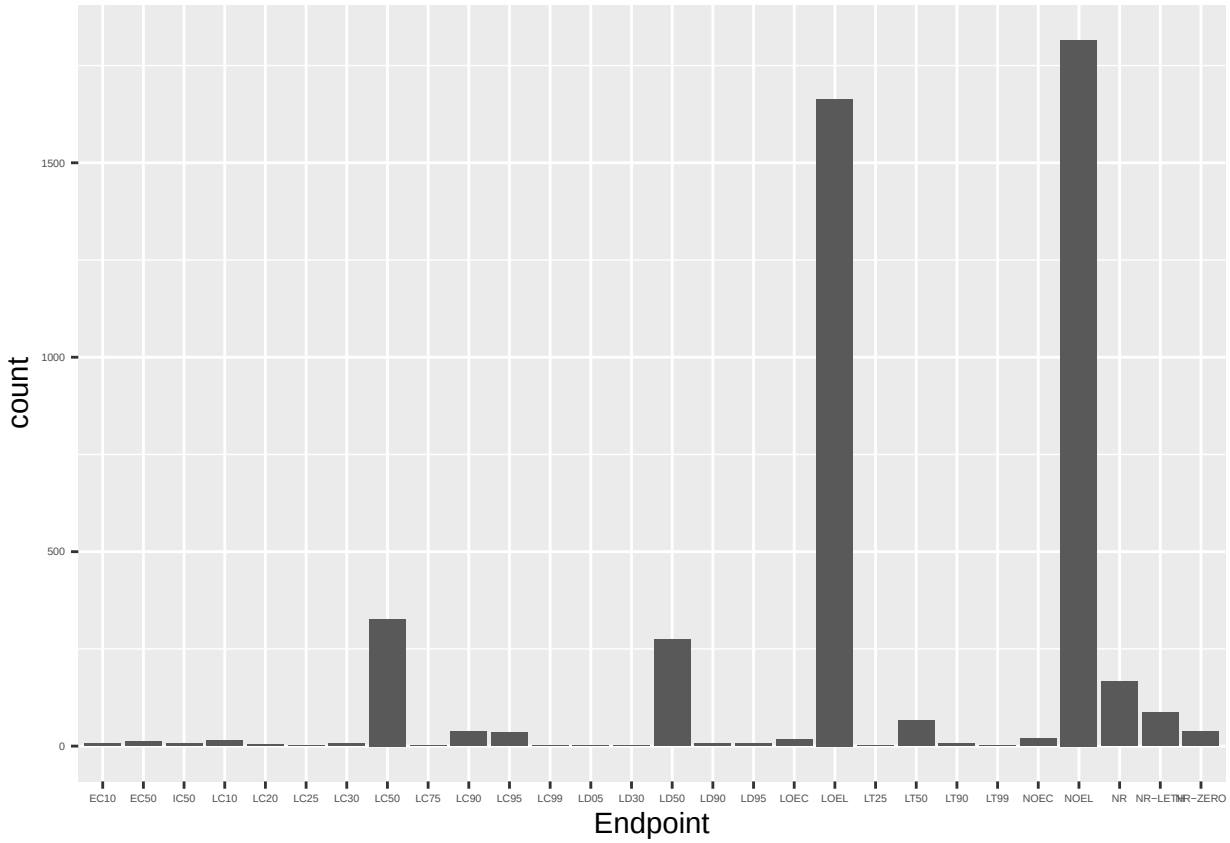
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The vast majority of test locations are in the lab or in a natural field setting. The number of tests in these two locations were relatively similar from 1990 until 2010. However, after 2010, the number field tests fell dramatically and lab tests became the dominant method.

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.

```
#Made a bar graph to show counts of various endpoints

ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar() +
  theme(axis.text=element_text(size=4))
```



Answer:The two most common Endpoints are LOEL and NOEL (In order to read these on the bar graph I had to reduce the size of the axis labels). LOEL is defined as the lowest concentration of a chemical producing effects that were significantly different than the effects of the controls. NOEL is defined as the highest concentration of a chemical producing effects that were NOT significantly different than the effects of the controls. Thus, LOEL and NOEL values should be pretty similar.

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.

    Answer: It is not a date, so I have changed the format from factor to date below. Litter was collected on 8/2/2018 and 8/30/2018.

```
#Changing format from factor to date
class(Litter$collectDate)
```

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

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

#I used this to double check all collect dates were in August 2018.
Litter.collect.August.2018 <-
  Litter %>%
  filter(month(collectDate) == 8) %>%
  filter(year(collectDate) == 2018)
```

```
unique(Litter.collect.August.2018$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`?

```
#Determining number of distinct plots at Niwot
unique(Litter$plotID)
```

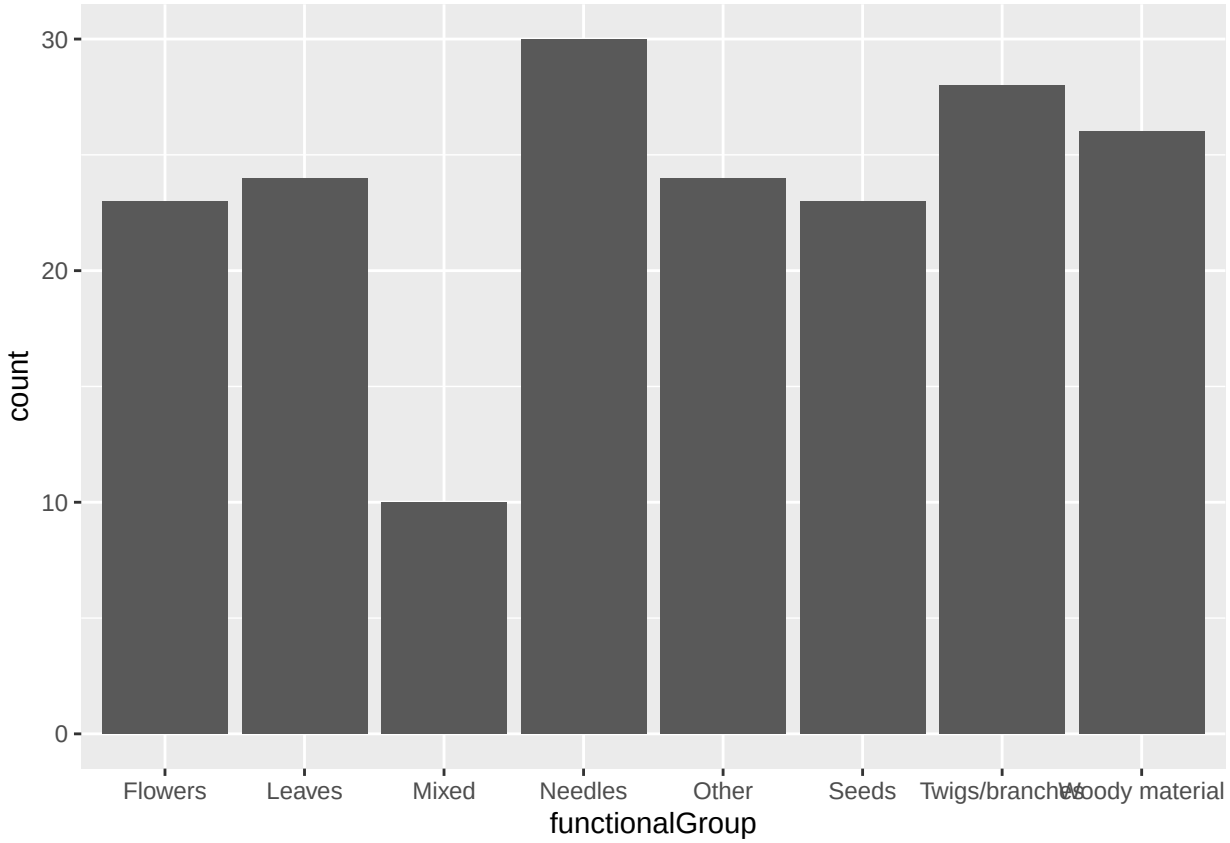
```
##  [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

Answer: There are 12 plots at Niwot Ridge. Unique is diffent than summary because unique gives you the number of unique values in a column. Summary, on the other hand, tells you the count for each unique value in a column. It does not provide the total number of unique values, which could become problematic for a large dataset.

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.

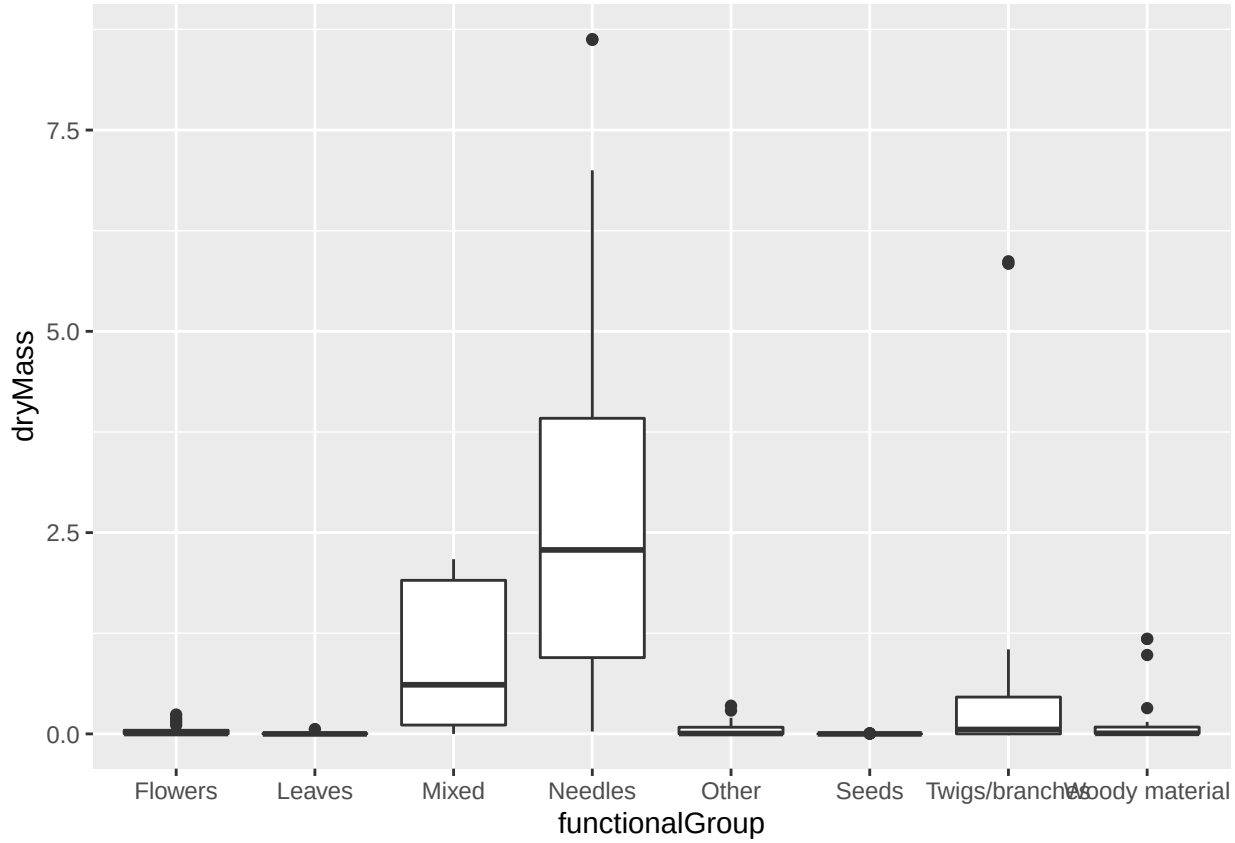
```
#Bar graph of litter types

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

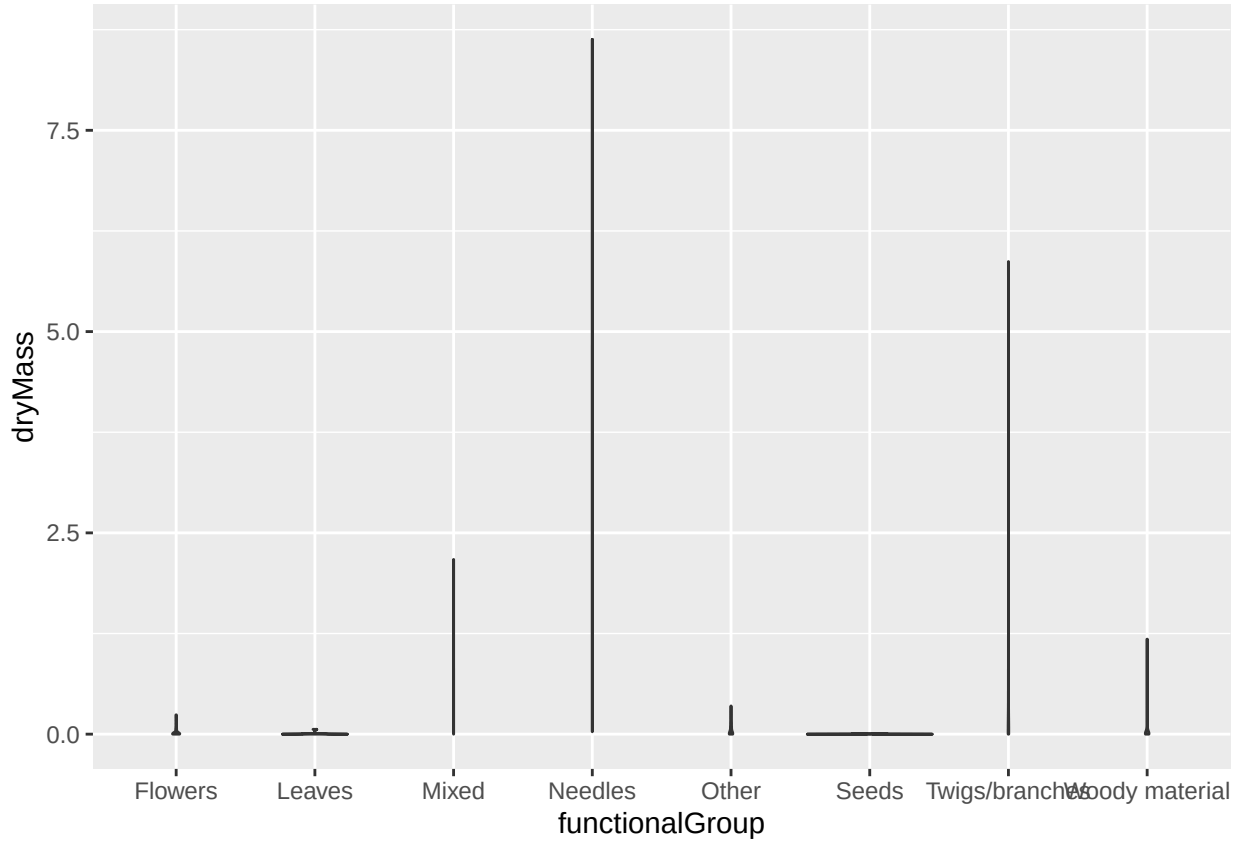


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#Created boxplot and violin plot of dryMass by functional group
ggplot(Litter, aes(functionalGroup, dryMass)) +
  geom_boxplot()
```




```
ggplot(Litter, aes(functionalGroup, dryMass)) +
  geom_violin()
```



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

> Answer: A boxplot is more effective because the distributions of dry masses are very different depending on the functional group. Leaves have many dryMass values that are nearly identical to or near zero, so the violin plot shows a horizontal line. Similarly, needles have dryMass values that are well spread out and do not form a bell curve. Thus, the violin plot shows a thin, vertical line. The box plot better represents all of the data.

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

> Answer: Needles, mixed, and twigs/branches have the highest dryMass values at Niwot Ridge.