Assignment 3: Data Exploration

Hanna Bliska

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.

```
#set and check working directory
setwd("~/Desktop/EDA-Fall2022")
getwd()
```

```
## [1] "/Users/hbliska/Desktop/EDA-Fall2022"
```

```
#install.packages(tidyverse)
library(tidyverse)
#neonics dataset
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = TRUE)
#litter dataset
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: We might be interested in the ecotoxicology of neonicotinoids on insects because many insects consume plant matter and rely on plants for pollen and nectar. If neonicotinoids are applied to agricultural products, it may be the case that this is having negative impacts on insects who are exposed to the insecticide when they forage or visit plants for pollintation. Studying the ecotoxicological effects of the insecticide on insects will help illucidate any negative impacts to insects, which could have consequences for ecosystems as pollinators such as bees play essential roles in helping agricultural and other ecosystems thrive.

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: We might be interested in studying litter and woody debris from the Niwot Ridge station in the Rocky Mountains of Colorado because litter and woody debris form the available fuel for wildfires. With warmer winters and increased year-round temperatures, litter and woody debris dry and are more easily ignited, starting wildfires. If we are able to measure abundances of litter and woody debris, we may be able to identify areas where fuel loads for wildfires are higher and monitor these areas for fire risk.

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 and woody debris sampling occurs within vegetation tower plots. Tower plots in dense forest are 40x40 meters while tower plots in lower-saturated vegetation are 40x40 meters and 20x20 meters. 2. Between one and four litter trap pairs are placed within each sampling plot, as a paired trap is placed every 400 meters squared. A litter trap pair consists of both an elevated trap and a ground trap. 3. Elevated traps are sampled temporally depending on the type of vegetation present. For the time period where deciduous leaves change color and shed, elevated traps in deciduous forests are sampled once every two weeks. At sites with evergreen trees, elevated traps are sampled once every one-to-two months. The network samples ground traps once per year.

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: The most common effects that are studied are population and mortality. These effects may be specifically of interest because they can serve as indicators for negative impacts on insects from neonictinoids. We could use statistical analysis to test if areas where neonictinoids have been applied produce significant negative impacts on mortality and population numbers for insects.

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: The six most commonly studied species in the dataset are honey bees, parasitic wasps, buff tailed bumblebees, Carniolan honey bees, bumble bees, and Italian honeybees. Honey bees and bumble bees are both vital pollinators of agricultural crops and flowers. Thus, they may be of particular interest to those concerned with the health of agricultural systems. Honey bees also produce honey, which is an important resource consumed by people in the United States. Parasitic wasps do not play a large role in pollination like the other five species listed do; however, parasitic wasps do have a role in agricultural ecosystems because they prey upon agricultural pests and are thus considered to be biological control agents (Chen et al. 2018). Therefore, farmers and other agricultural stakeholders may be concerned with the health of parasitic wasp populations due to the important role they place in aiding agricultural production.

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.)
```

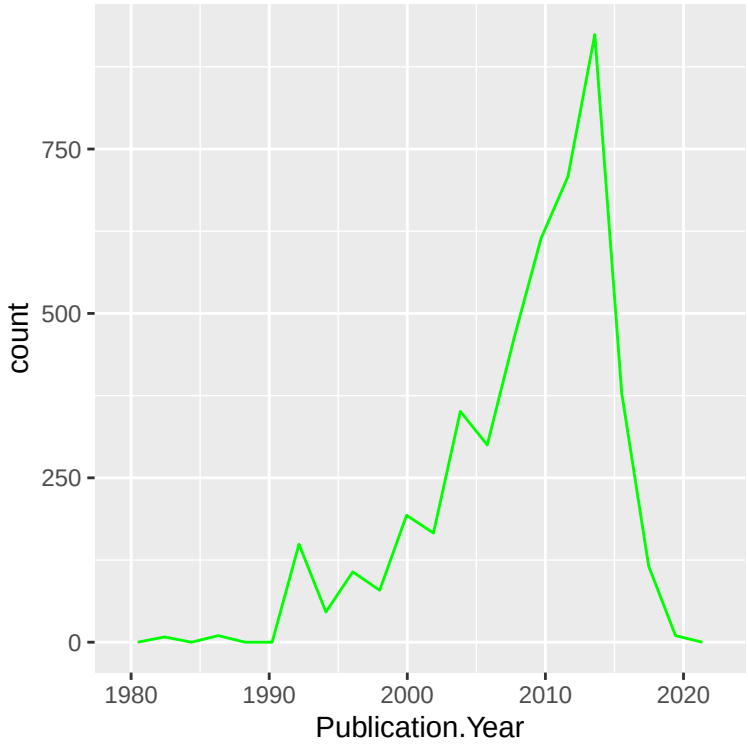
```
## [1] "factor"
```

Answer: The class of Conc.1..Author is factor. The reason that R is recognizing this column as a factor column is because the data in this column are not continuous, numerical values. Rather, there are ranges of data (e.g., <1.5) and approximations (e.g., ~41) recorded. These values cannot be considered continuous, rather, they are categorical observations. I hypothesize that some of the concentrations were recorded in this manner due to the difficulty of detecting concentrations of chemicals from a laboratory science perspective; it may be the case that the equipment used could not detect a precise numeric value from a given sample, and thus ranges or approximations were recorded.

Explore your data graphically (Neonics)

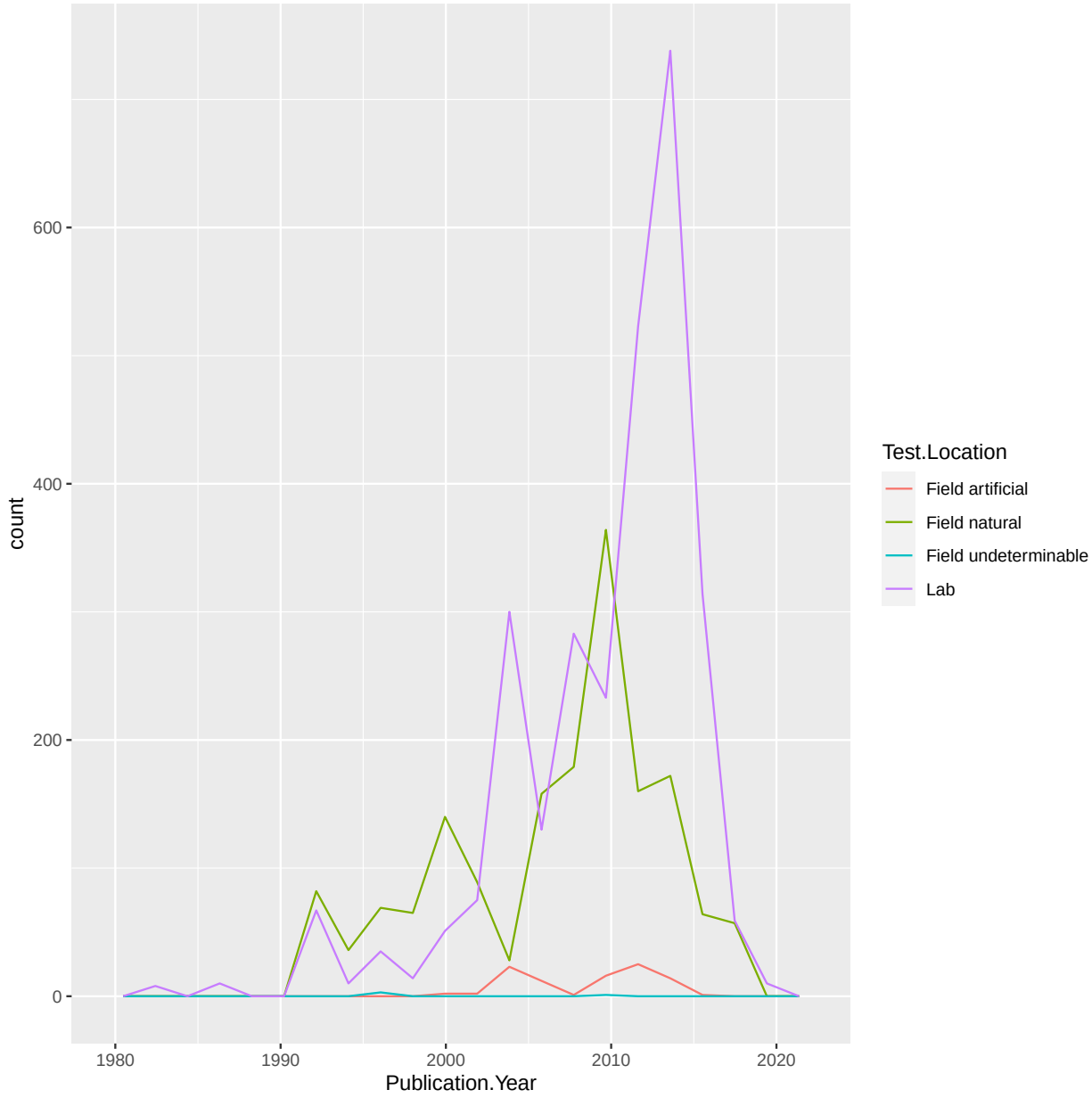
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

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



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)
```



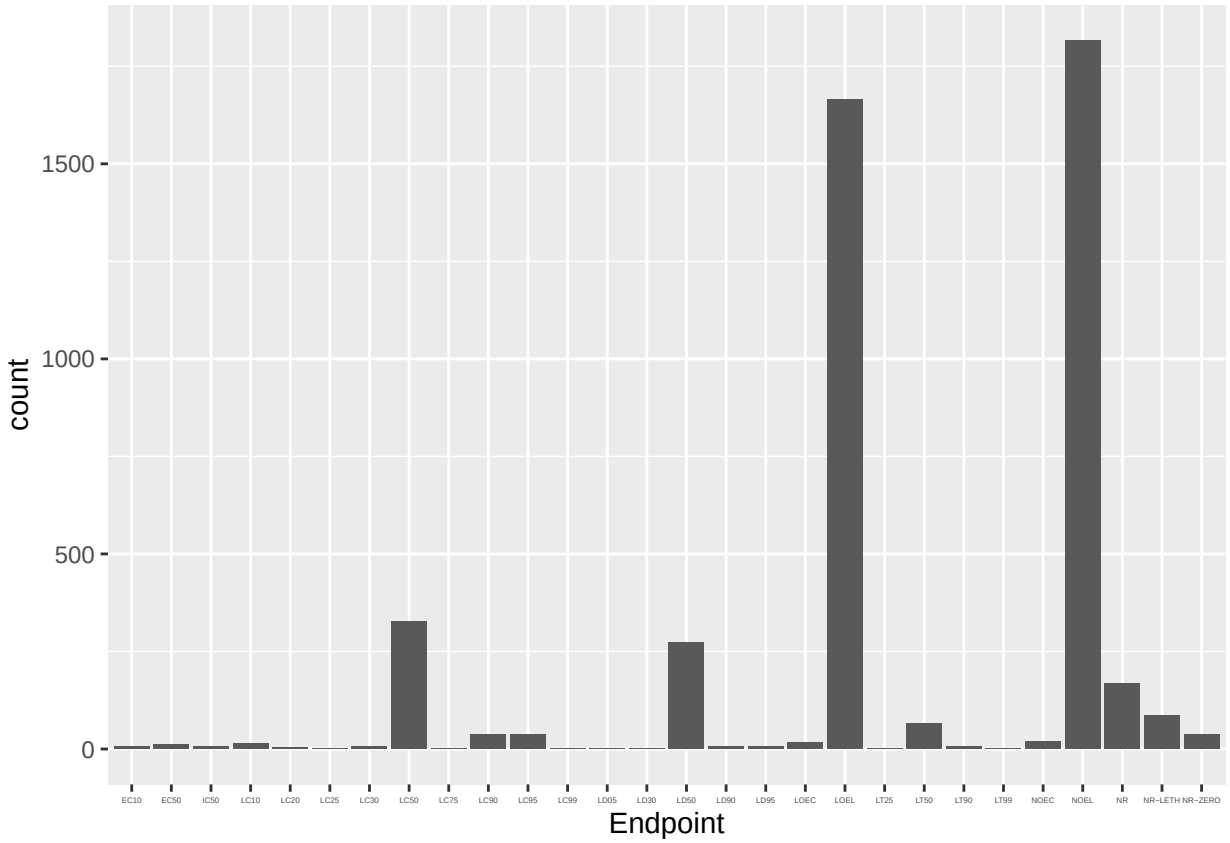
Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: The most common test locations are lab settings and natural field studies. The graph indicates that lab settings became much more common test locations for studies in our data set after 2010, reaching a maximum around 2014 with approximately 550 publications. Natural field studies, in comparison, were most common in studies in our dataset prior to 2010, reaching a maximum of approximately 350 studies between 2009-2010. This graph suggests that the studies referenced in our dataset reflect trends in test locations for exotoxicological studies, with lab settings becoming more common in recent years. However, after 2015, both lab settings and natural field studies exhibited a sharp decline in our dataset; this could suggest that we have less observations in our dataset of sources from 2015-2020, or that these test locations are becoming less utilized in the literature.

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, aes(x = Endpoint)) +
  geom_bar() +
  theme(axis.text.x = element_text(size = 3))
```



Answer: The two most common endpoints are LOEL and NOEL. According to ECOTOX_CodeAppendix, LOEL is defined for terrestrial samples as the lowest-observable effect level, meaning it is the lowest concentration or dose of chemicals that produced effects that were significantly different from reponses of control samples. NOEL is defined for terrestrial samples as the no-observable effect level, meaning it is the highest concentration or dose of chemicals that produces effects not statistically significantly different from responses of control samples.

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.

```
#Check class
class(Litter$collectDate)
```

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

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

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

```
#Determine dates of litter 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`?

```
#Use unique function to determine # and ID of unique plots
unique(Litter$plotID)
```

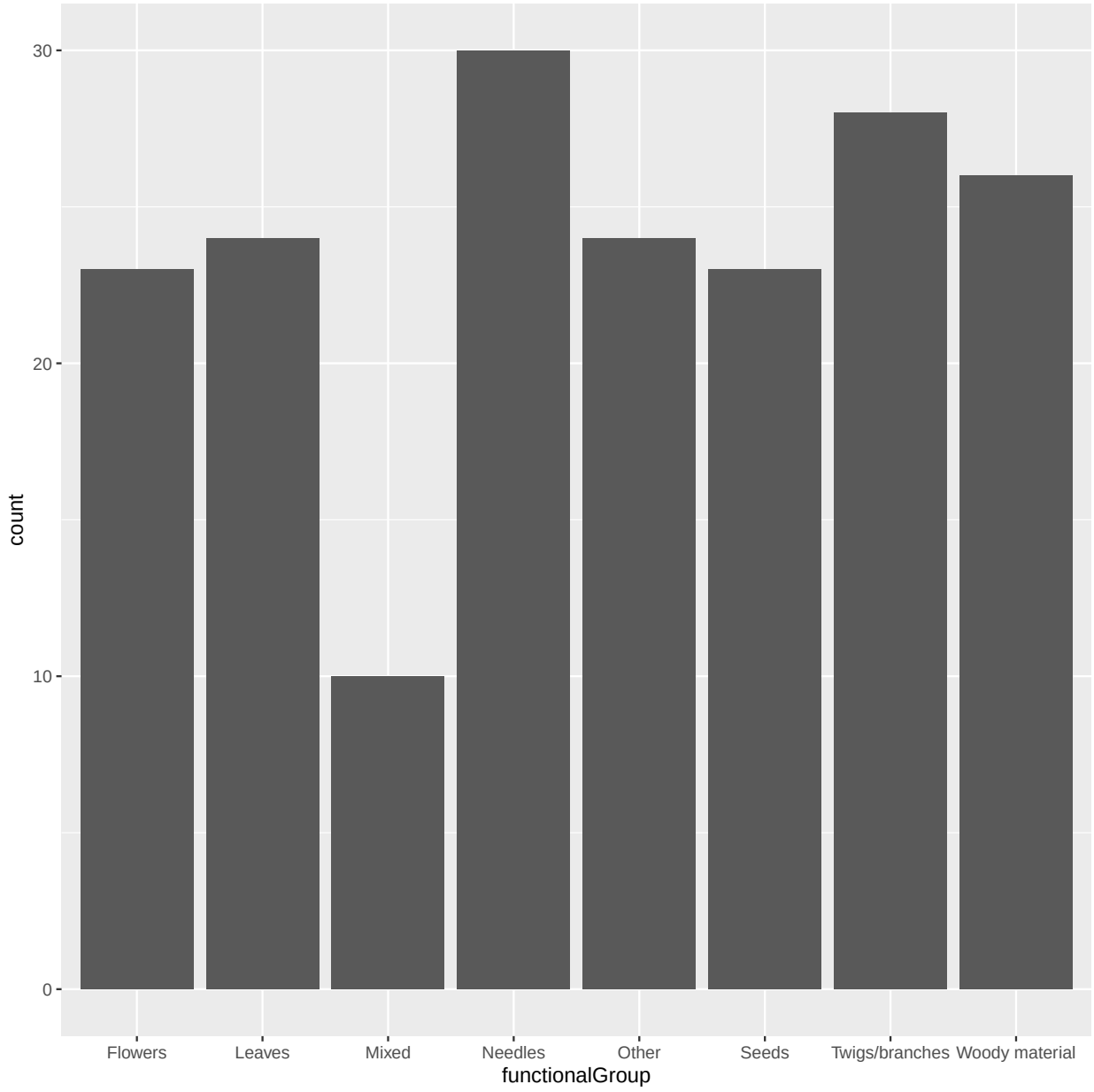
```
##  [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

```
#Use summary function to determine # of times a plot was sampled
#summary(Litter$plotID)
```

Answer: 12 unique plots were sampled at Niwot Ridge. `Summary` reports the frequency of samples taken at each plot (for example, Plot NIWO_040 was sampled 20 times.) `Unique` reports the number of unique plots and their IDs, which allows the user to quickly determine unique values without reading information pertaining to duplicates.

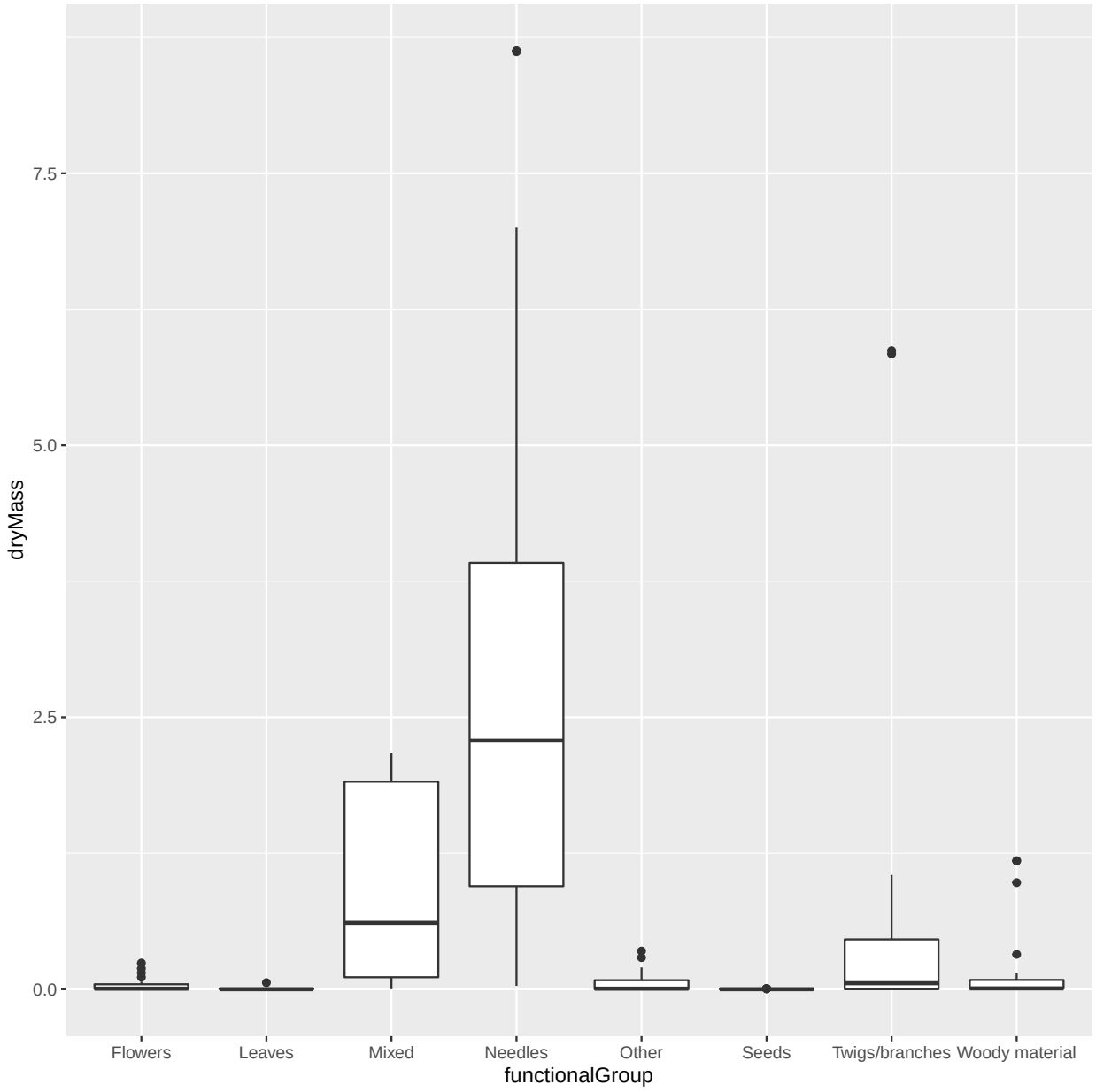
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

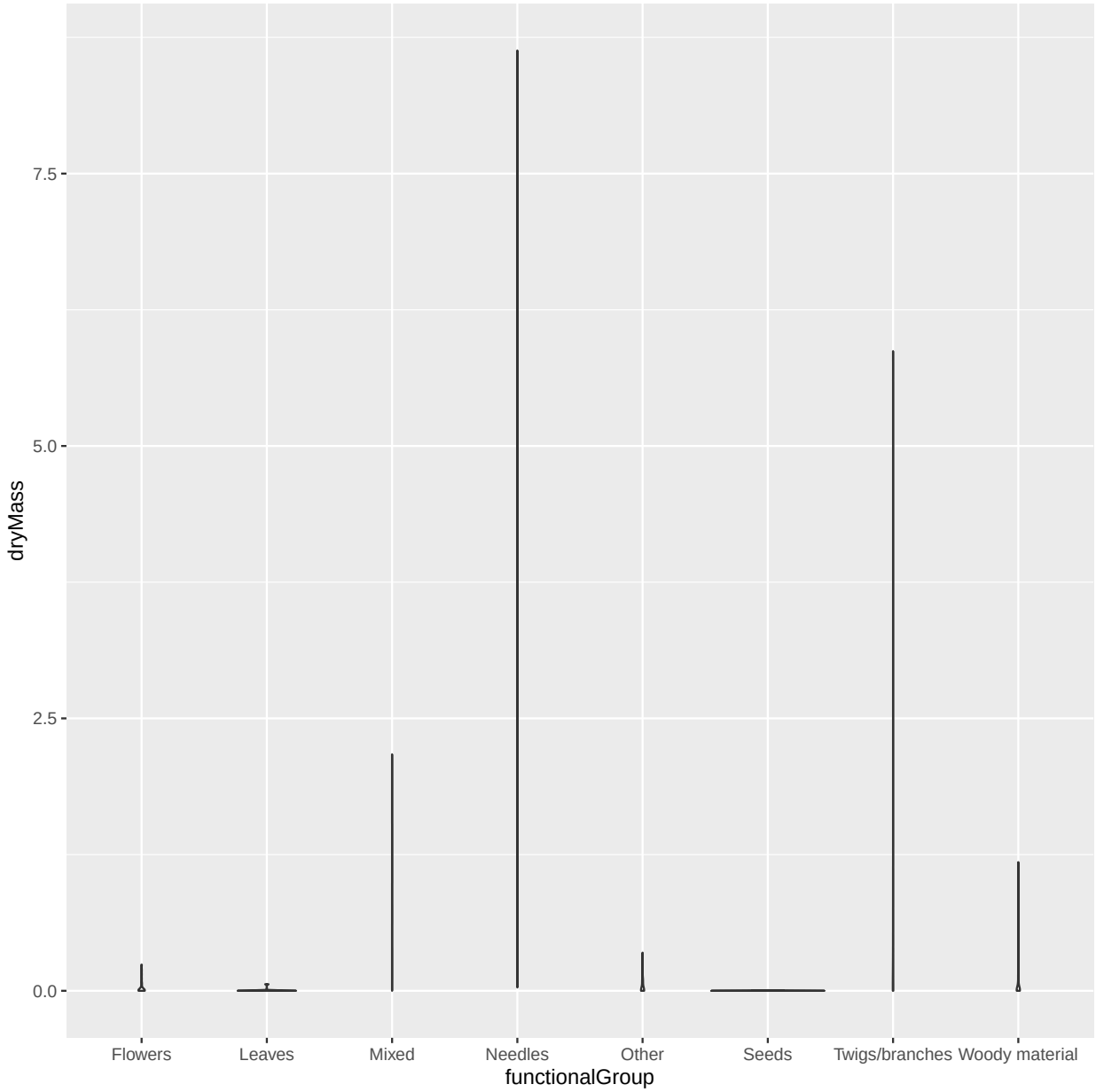


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#Creating a boxplot
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass))
```




```
#Creating a violin plot
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass))
```



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

> Answer: The boxplot produced here effectively visualizes important summary statistics (e.g., the median and interquartile range) for this dataset. A violin plot is not helpful for visualizing these data because the functional groups either have very long distributions for measurements of dry mass or very short distributions. Violin plots are most useful for visualizing patterns in distributions of the data; for example, if many of the data were clustered around the median, or minima/maxima. Since this is not the case, the violin plots produce straight long and short lines, and no useful information is gleaned from the chart.

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

> Answer: Needles and mixed litter tend to have the highest biomass at these sites.

References: Chen L, Lang K, Bi S, Luo J, Liu F, Ye X, Xu J, He K, Li F, Ye G, Chen X. WaspBase: a genomic resource for the interactions among parasitic wasps, insect hosts and plants. Database (Oxford). 2018 Jan 1;2018:1-9. doi: 10.1093/database/bay081.