Elise Boos, Section #1

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.**

```
# set wd
setwd("~/Desktop/Data_Analytics/Environmental_Data_Analytics_2022")
# load packages
require(tidyverse)
# load 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: neonicotinoid is a common insecticide used in agriculture due to its water soluability and alleged low-toxity to insects, though this claim has come into question. It has been seen to have a negative effect on pollinators like bees and also potentially detrimental to aquatic health as well. Its important to study the effect of neonicotinoids on insects because these types of insecticides are so widely used.

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 to ecosystems and are of interest because they play a role in carbon budgets and nutriet cycling as well as are energy sources for decomposers and microorganisms among other things. Litter and woody debris also influence aquatic ecosystems by influencing water flow dynamics and diversifying habitat for organisms.

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: * Litter and fine woody debris were collected using elevated and ground traps respectively. * Trap placement within plots were targeted or randomized, depending on the vegetation. Sites with $>$ 50% aerial cover of woody vegetation $>$2m in height, placement of litter traps is random. In sites with $<$ 50% cover of woody vegetation trap placement is targeted such that only areas beneath qualifying vegetation are considered for trap placement. * Temoral sampling varies with vegetation site. For example ground traps were sampled once per year and sampling was done every two weeks for elevated traps in decidous forest sites.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
# dimension function
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 on only effect column
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 and mortality. These effects are of interest because they have large implications. Mortality and changes to population dynamics have the potential to alter not only species, but entire ecosystems. In bees this could indicate collany collapse which would have cascading effects.

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.

```
# show only the first six (meaning six highest) of the
# species common name column summary
head(summary(Neonics$Species.Common.Name))
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee      Italian Honeybee
##                   152                   140                   113
```

Answer: Honey Bee, Parasitic Wasp, Bugg Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, Italian Honeybee. All of these species are types of bees or wasps. These are of interest because they are important species for pollination and are not harmful, but helpful to crops.

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

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

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

```
# visualize data
head(summary(Neonics$Conc.1..Author.))
```

```
## 0.37/   10/   NR/    NR     1  1023
##   208   127   108    94    82    80
```

Answer: The class is as a factor. This is because the values in this column are not just numbers. There contains rows with slashes after the numbers as well as 'NR' values.

Explore your data graphically (Neonics)

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

```
# plot of number of studies per year in duke blue
ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year), bins = 50,
    color = "#00009c") + labs(x = "Publication Year", y = "Number")
```

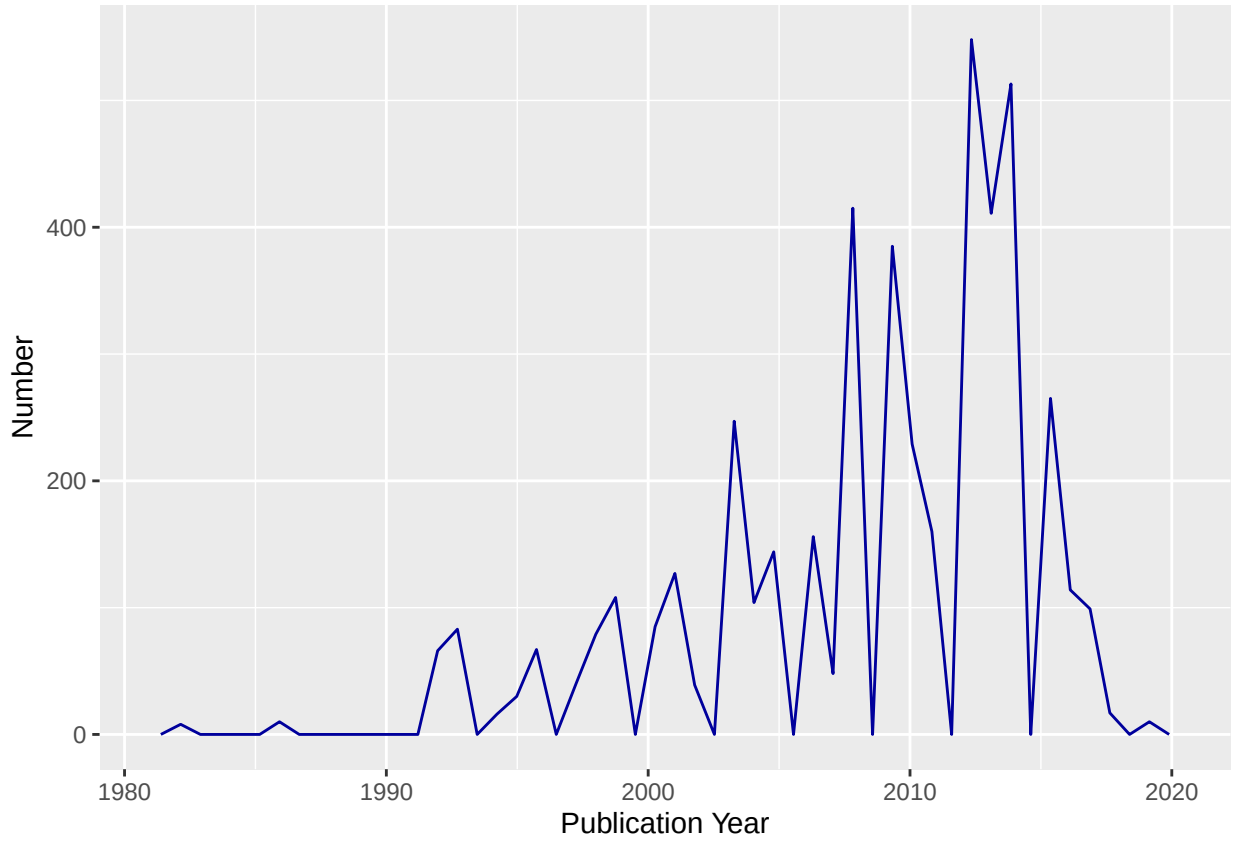


Figure 1: Number of studies conducted by publication year

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

```
# plot of publication year but classified by test location
ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year, color = Test.Location),
    bins = 50) + labs(x = "Publication Year", y = "Number") +
    scale_color_discrete(name = "Test Location", labels = c("Artificial Field",
        "Natural Field", "Undeterminable Field", "Lab"))
```

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

> Answer: The most common test locations are in a lab and a natural field. There are fluctuations between lab and natural field on which are the most common test locations. There is a big spike in lab locations between 2010 and 2015.

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.

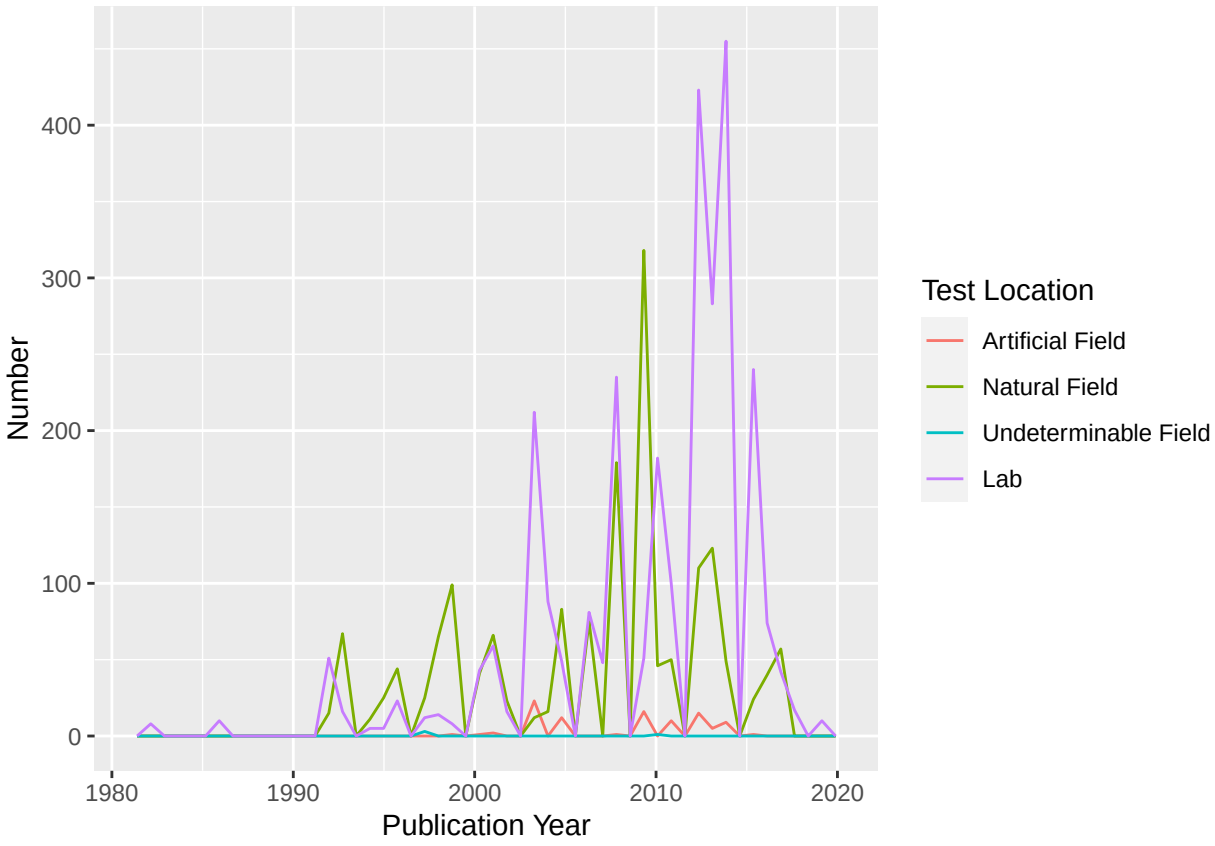
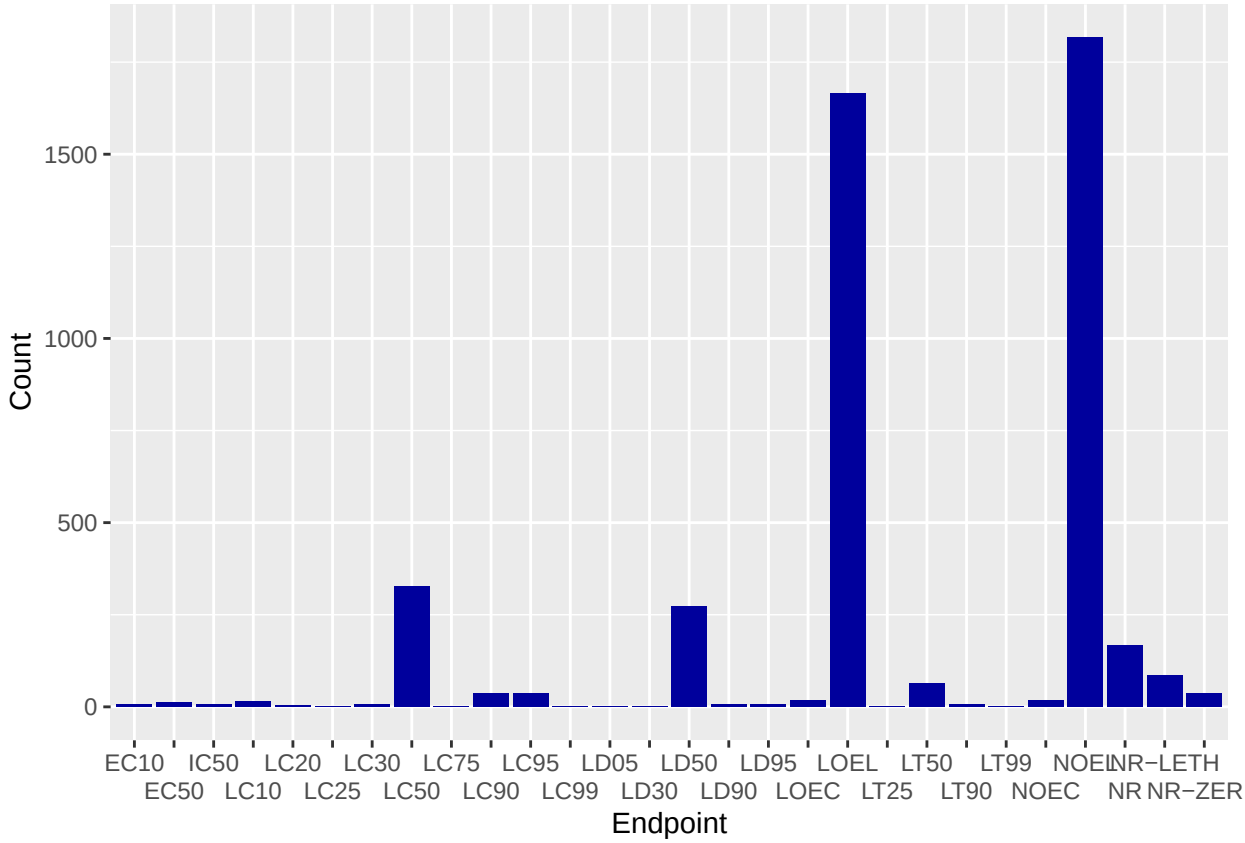


Figure 2: Number of studies conducted in each test location by publication year

```
# bar graph of endpoints in duke blue
ggplot(Neonics, aes(x = Endpoint)) + geom_bar(fill = "#00009c") +
    labs(x = "Endpoint", y = "Count") + guides(x = guide_axis(n.dodge = 2))
```



Answer: Two most common endpoints are LOEL and NOEL. LOEL = Lowest-observable-effect-level meaning lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls and NOEL = No-observable-effect-level meaning highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test.

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.

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

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

```
# set as date
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
# check still written right
head(Litter$collectDate)
```

```
## [1] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
## [6] "2018-08-02"
```

```
# check if class now date
class(Litter$collectDate)
```

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

```
# find unique dates, filter out repeats
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`?

```
# run unique function of plotID
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
```

```
# compare to summary function
summary(Litter$plotID)
```

```
## 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: Unique tells you what values in the column are unique (aka filters out any repeat data). Unique tells us the number of unique values as well, in this case 12. Summary tells you the amount the unique value appears in the coloumn.

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(fill = "#00009c") +
    labs(x = "Litter Type", y = "Count") + guides(x = guide_axis(n.dodge = 2))
```

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

```
# boxplot
ggplot(Litter) + geom_boxplot(aes(x = functionalGroup, y = dryMass),
    fill = "spring green 4") + labs(x = "Litter Type", y = "Dry Mass") +
    guides(x = guide_axis(n.dodge = 2))
```

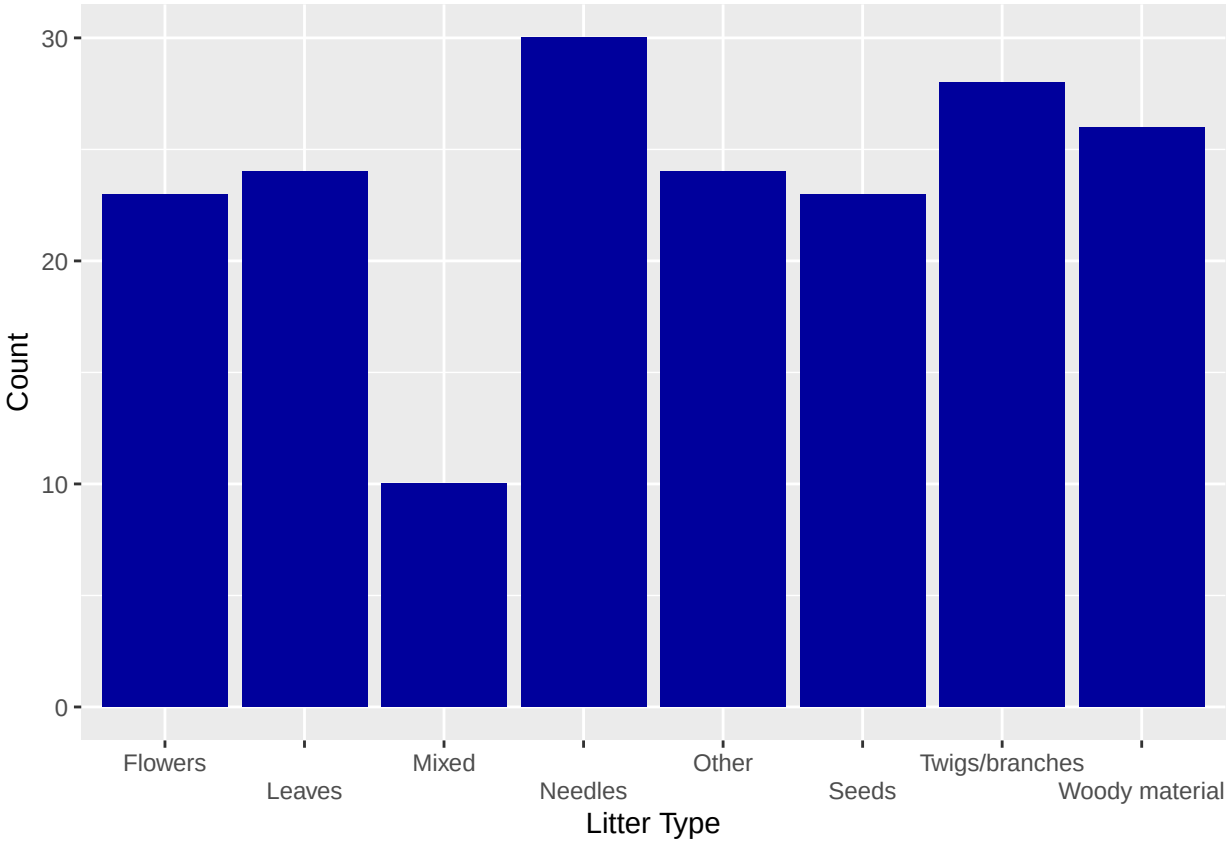


Figure 3: Types of litter collected at Niwot Ridge sites

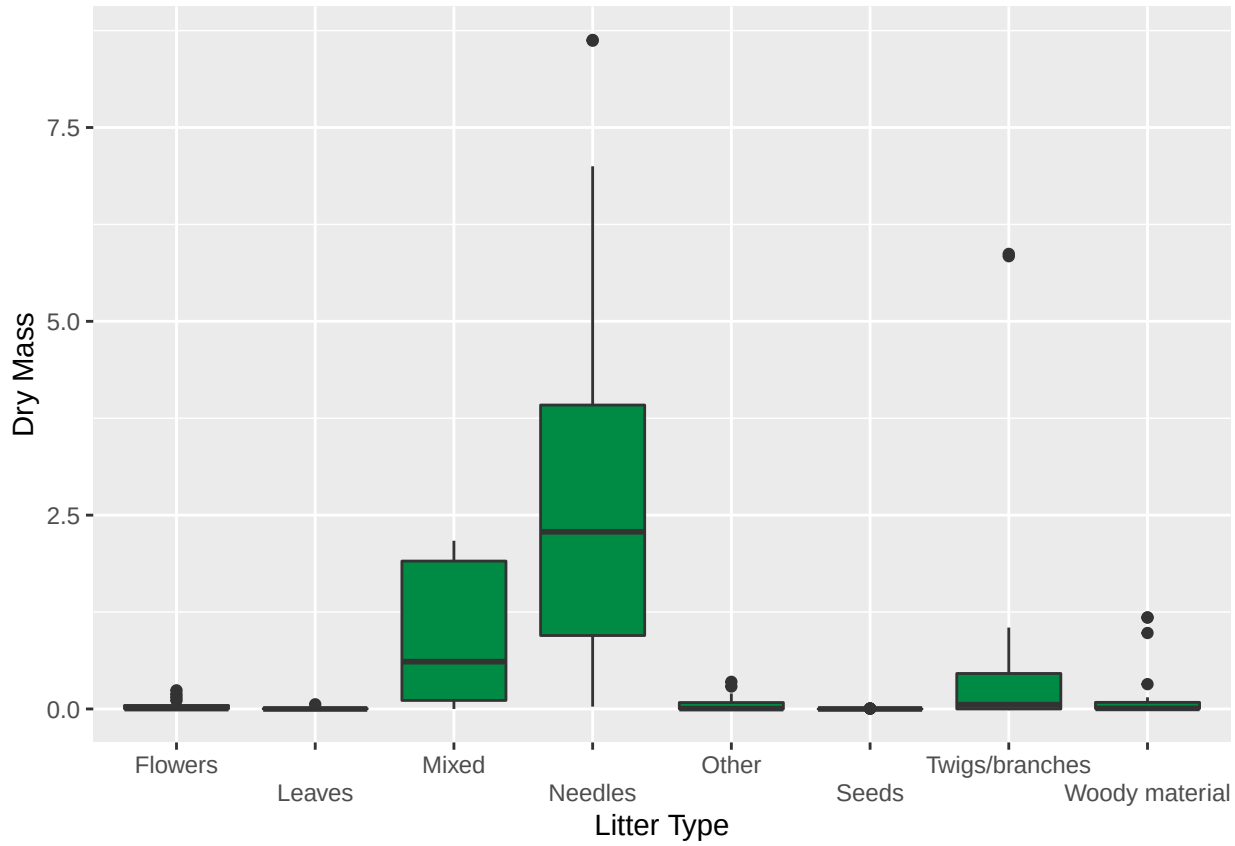


Figure 4: Dry mass of litter types collected at Niwot Ridge sites

```
# violin plot
ggplot(Litter) + geom_violin(aes(x = functionalGroup, y = dryMass),
    draw_quantiles = c(0.25, 0.5, 0.75)) + labs(x = "Litter Type",
    y = "Dry Mass") + guides(x = guide_axis(n.dodge = 2))
```

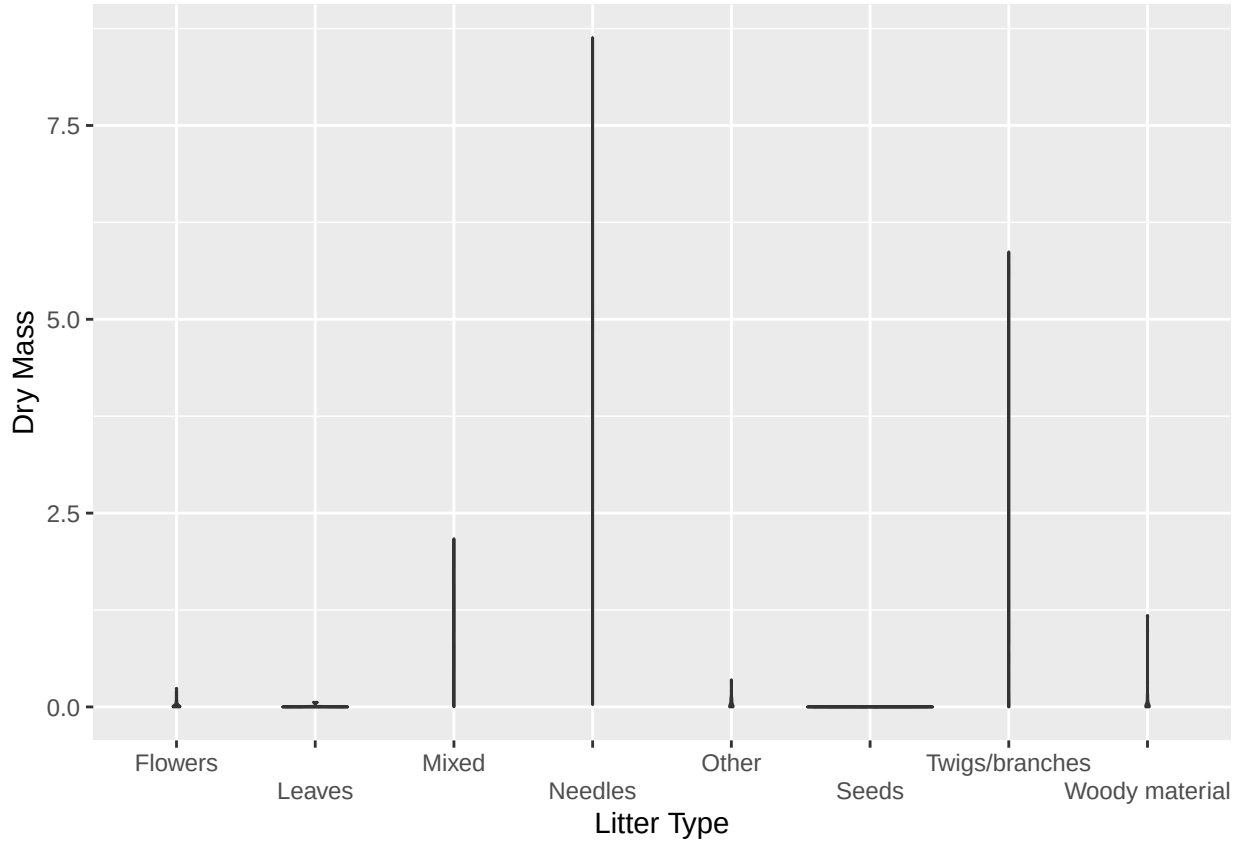


Figure 5: Dry mass of litter types collected at Niwot Ridge sites

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

> Answer: The boxplot is more effective because violin plots show the density in the data, but in this case the data is not densly clustered and more evenly distributed resulting in flat lines. In this case summary statistics suffice and a box plot visualizes the data better.

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

> Answer: Needles has the highest then mixed litter.