Assignment 3: Data Exploration

Langston Alexander, Section #4

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.**

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
Neonics <- read.csv("C:/Users/lwa8/Documents/R/ENV872/Environmental_Data_Analytics_2022/Data/Raw/ECOTOX

Litter <- read.csv("C:/Users/lwa8/Documents/R/ENV872/Environmental_Data_Analytics_2022/Data/Raw/NEON_NI

head(Neonics)
head(Litter)
```

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: Insects are a vital part of any ecosystem. The health of insect populations directly impact other species in the ecosystem either directly or indirectly. It is important to understand how these widespread insecticides effect and are transfered by both the intended and unintended targets. By unintended target I mean that once these insecticides are applied to crops they have the potential to effect any insect that comes in contact with the crop, whether that insect is a pest or not. The effect on these non-pest insects, like pollinators, could have consquences on the ecosystem that outweigh the benefits of pest deterence.

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: A major reason woody debris and litter are important to understand, especially in forests located in the Western USA, is they are signficant contributors to wildfire intensity. Woody debris and forest litter act as fuel for wildfire. When debris and litter build up it adds more fuel for wildfires to burn, increasing their intensity. Understanding how forest litter and debris accumulates and the amount carried by healthy forests could help forest managers plan for wildfire mitigation in their domains.

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 is defined as material with a butt end diameter of <2cm and a length of <50cm. It is collected in elevated traps. Fine wood debris is material with a butt end diameter of <2cm and a length of >50cm. It is collected in ground traps. *Ground traps are sampled once per year. Elevated traps are sampled either once every 2 weeks or once every 1-2 months depending if sites are populated by deciduous or evergreen trees.* One elevated and one ground trap is deployed for every 400mˆ2 plot area, resulting in 1-4 trap pairs per plot.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) #4623 rows x 30 columns
```

```
## [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: By far the most studied effects of neonicotinoids on insects are mortality and population with 1493 and 1803 studies for each, respectively. The next two most commonly studied effects are behavior and feeding behavior with 360 and 255 studies respectively. Mortality and population effects of neonicotinoids may be the most commonly studied for two reasons: 1. When testing if neonicotinoids negatively effect certain insects the initial studies may want to see if the insecticides kill the test subjects and if those deaths have an impact on overall population. More complex studies, like those looking at change in behavior may follow these initial studies. 2. It may be easiest to study mortality and population 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.

```
species_summary <- summary(Neonics$Species.Common.Name)
head(species_summary)
```

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

Answer: From initial look at the most commonly studied species it seems they are all types of bees, with the exception of the parasitic wasp. I would assume that all of these insects are common pollinators as well, which explains the interest in studying the effects of neonicotinoids on them. Pollination is an important and fundamental ecosystem service which draws the attention of lots of researchers.

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.)
```

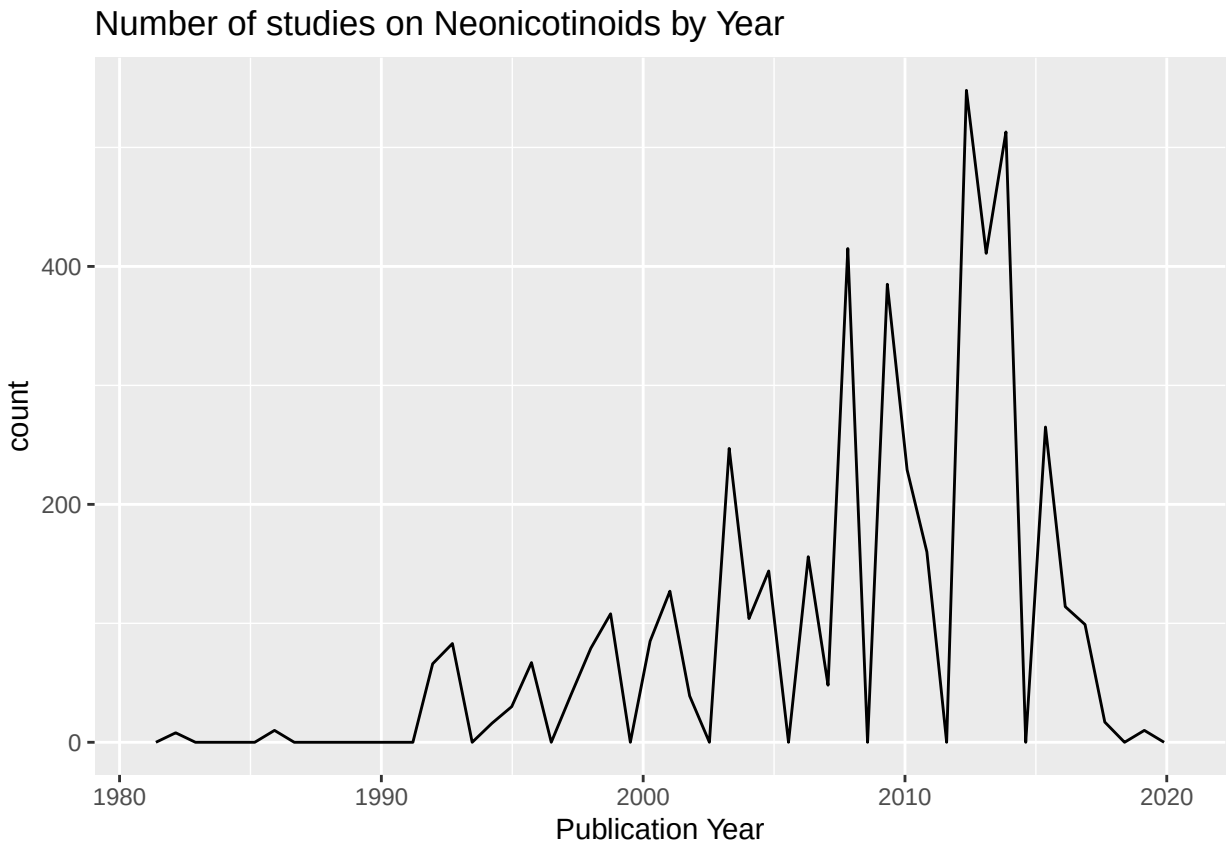
```
## [1] "factor"
```

Answer: Conc.1..Author is categorized as a factor in the dataset. It isn't numeric because several of the data points in the variable are characters, including "NR" and concentration data followed by a "/". Each variable can only contain one datatype so when R finds characters in a variable it switches all data to characters.

Explore your data graphically (Neonics)

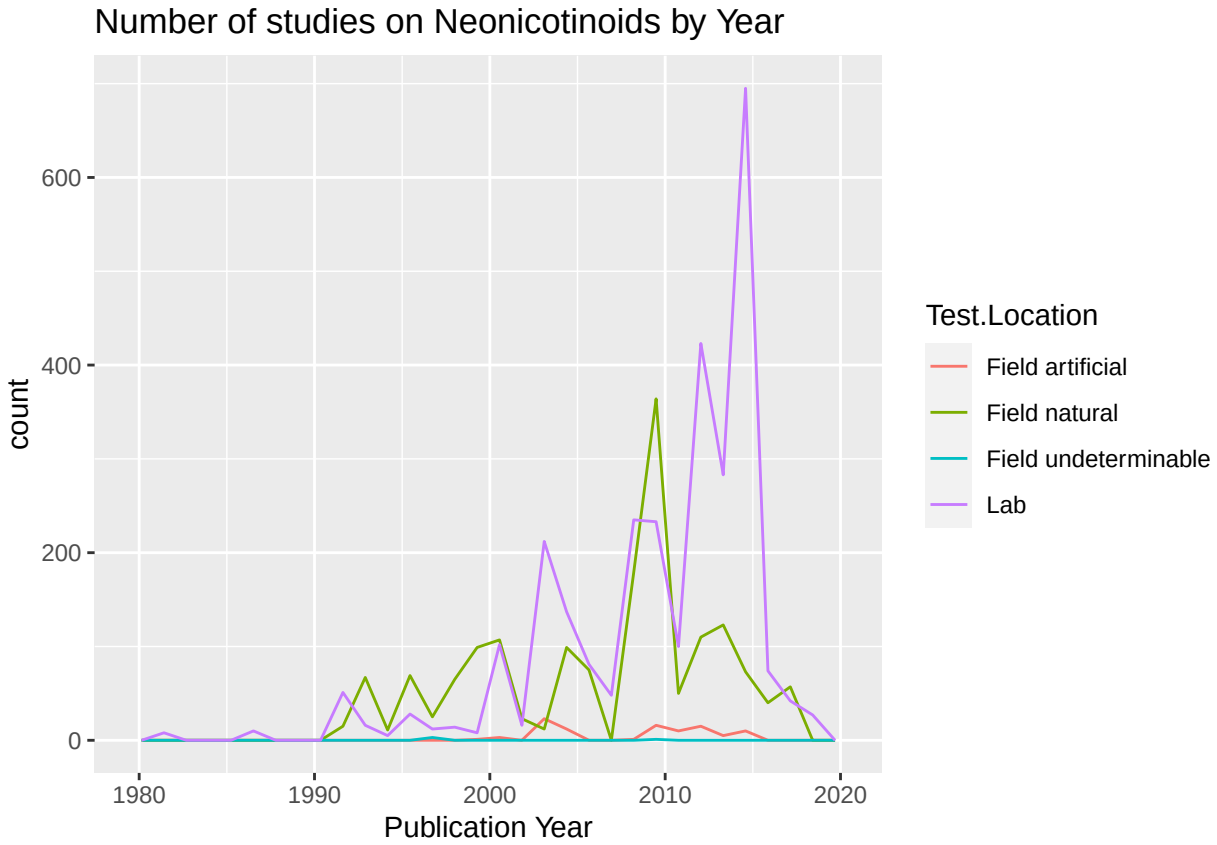
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics)+
  geom_freqpoly(aes(x= Publication.Year), bins=50)+
  labs(x= "Publication Year", title = "Number of studies on Neonicotinoids by Year")
```



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=30)+
  labs(x= "Publication Year", title = "Number of studies on Neonicotinoids by Year")
```

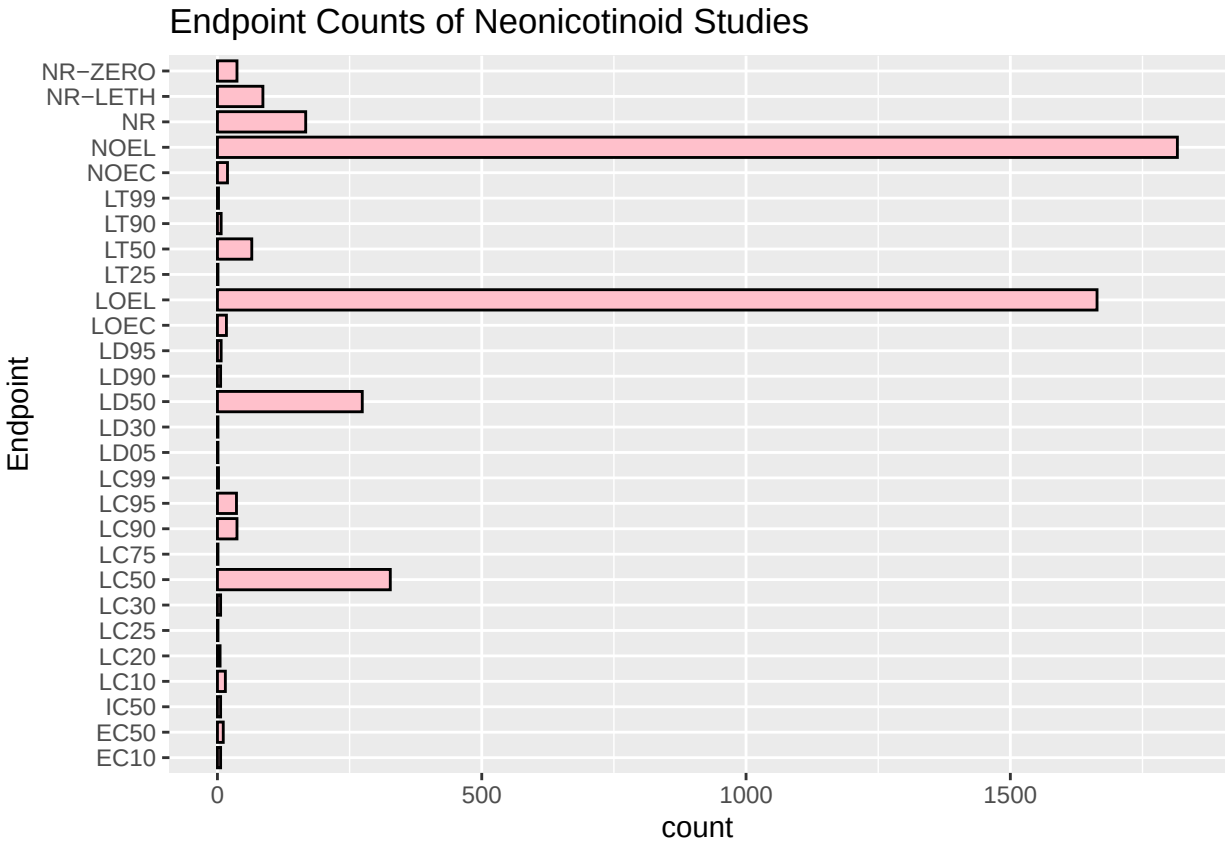


Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: Up until 2010 the amount of testing in natural field locations and in a lab were comparable. Since 2010, testing in a lab environment has increased significantly while testing in natural field locations has declined. Testing in artificial field locations and data labled undeterminable have been negligible since 1980.

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(fill = "Pink", color = "Black", width = .8)+
  coord_flip()+
  labs(title = "Endpoint Counts of Neonicotinoid Studies" )
```



Answer: The two most common endpoints are NOEL and LOEL. NOEL stands for no-observable-effect-level. NOEL means that the highest dose of the neonicotinoids did not produce a result different from the control group. LOEL stands for lowest-observable-effect-level. LOEL means that the lowest dose of neonicotinoids produced a result different from the control group.

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)

class(Litter$collectDate)
```

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

```
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`?

```
plotLevels <- unique(Litter$plotID)

nlevels(plotLevels)
```

```
## [1] 12
```

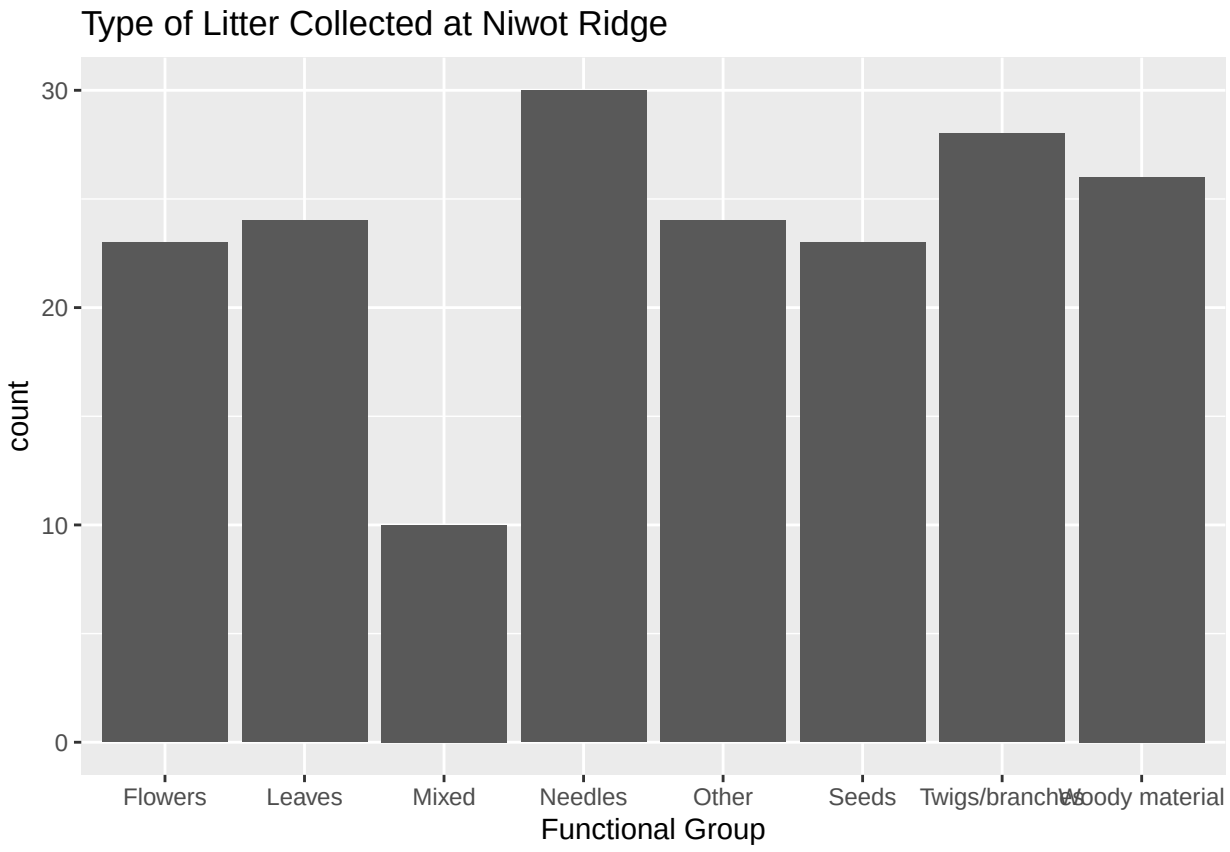
```
summary(Litter$plotID)
```

```
## 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: The unique function simply lists out the names of the unique levels of plots sampled at Niwot Ridge. Summary returns the names of unique plots along with the number of its instances in the variable.

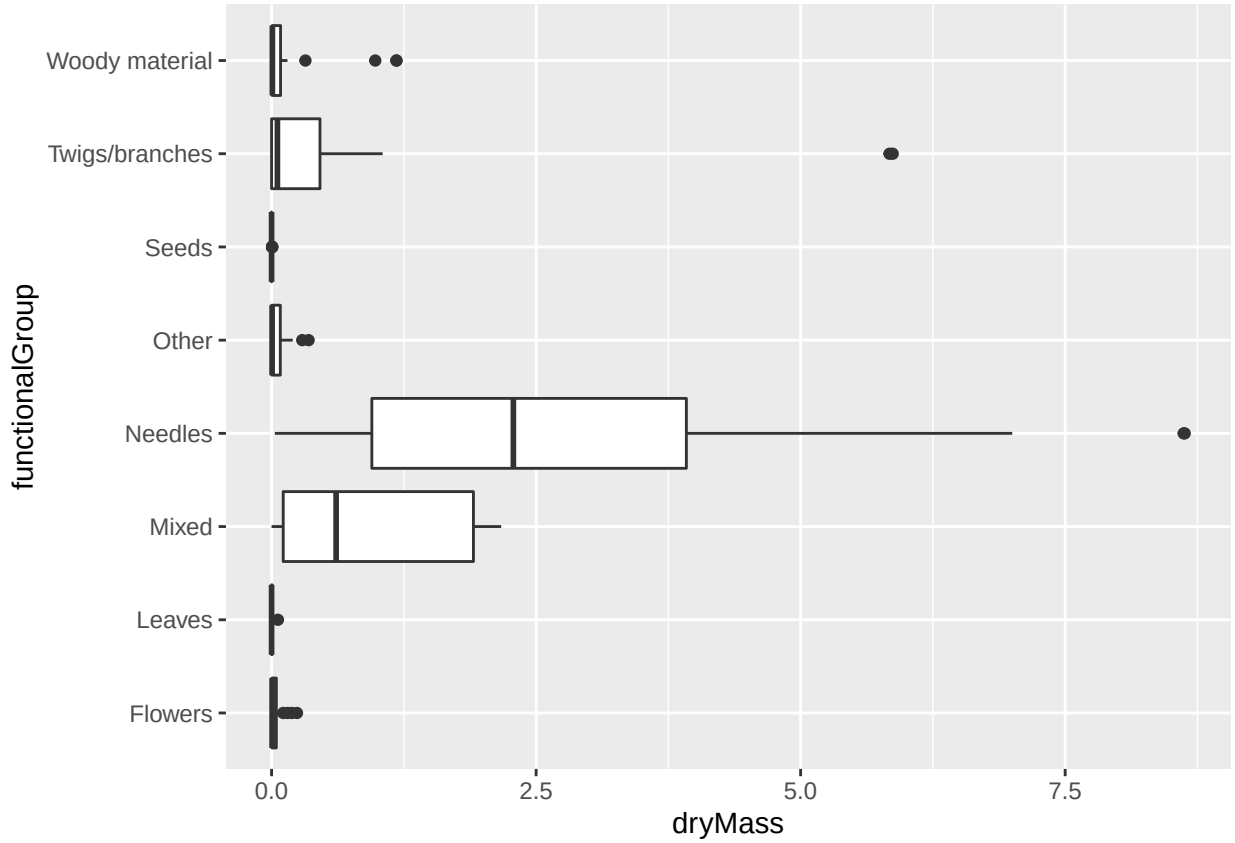
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(x=functionalGroup)) +
  geom_bar()+
  labs(x = "Functional Group", title = "Type of Litter Collected at Niwot Ridge")
```



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))
```

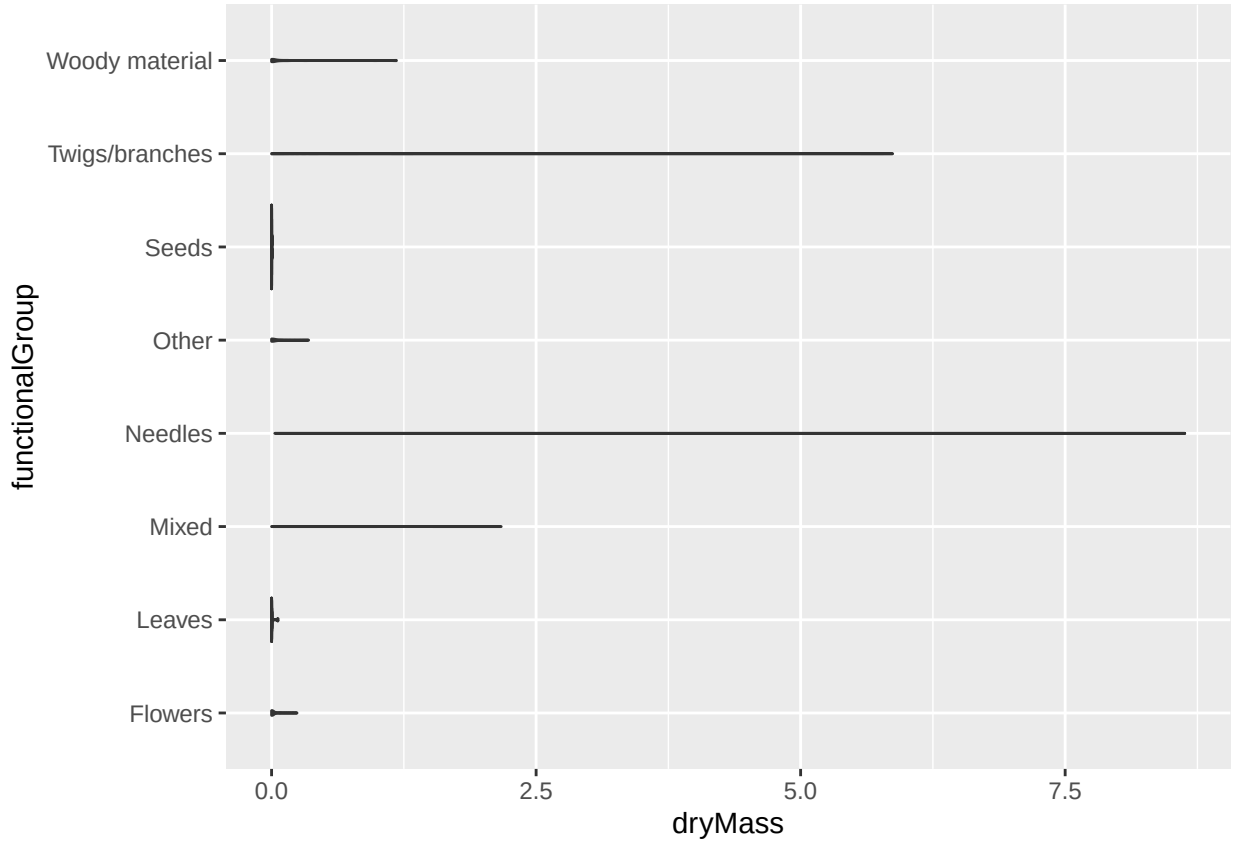



```
ggplot(Litter) +
  geom_violin(aes(x = dryMass, y = functionalGroup),  draw_quantiles = c(0.25, 0.5, 0.75))
```

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```



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

> Answer: Violin plots show the density of values laid over a box plot. In the dryMass variable, very few values, except for 0, are repeated more than once since dry mass is measured out to 3 decimal points. The violin plots come out as flat lines since the density for nearly every value is 1. This visualization provides very little useful information.

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

> Answer: Needles have the highest biomass at the sites with a median dry mass just under 2.5. Mixed litter has the next highes with a median dry mass around 1.