Assignment 3: Data Exploration

Claire Mullaney

OVERVIEW

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

Directions

1. Change "Student Name" on line 3 (above) with your name.
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., "Salk_A03_DataExploration.Rmd") prior to submission.

The completed exercise is due on Tuesday, January 28 at 1:00 pm.

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.

```
#Check working directory
getwd()
```

```
## [1] "/Users/clairemullaney/Desktop/ENV 872/Environmental_Data_Analytics_2020"
```

```
#Load necessary packages
library(tidyverse)

#Load datasets
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv")

Litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv")
```

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: Some insects play important ecosystem roles and perform tasks cruicial to human health and survival – for example, they pollinate flowers and crops. If neonicotinoids are possibly harming beneficial insects and preventing them from effectively completing these tasks, the impact of neonicotinoids on insect health would need to be studied. From a quick search, it appears that there is evidence that neonicotinoids are contributing to declines in bee populations, which would certainly cause scientists to devote resources to this area of study.

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: Litter and woody debris that fall to the ground in forests could be used to assess different element and chemical concentrations in the foliage. These concentrations could possibly be indicators of overall forest health, and anomolies or trends could be indicative of an event that caused the litter to contain more of a specific compound or element (e.g., deposition of mercury from anthropogenic emissions). These debris could also be used to study ecological productivity, nutrient and carbon cycling, and soil fertility.

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 were collected using both elevated and ground traps. Elevated traps primarily catch litter, and each is a 0.5 m^2 square with a mesh basket elevated about 80 cm above the ground. Ground traps primarily catch woody debris, and each is 3 m x 0.5 m.
   2. Sampling occurs only in plots near towers. If the area near the tower is forested, litter sampling takes place in 20 40m x 40m plots. If the sites near the tower have low-statured vegetation, litter sampling occurs in 4 40m x 40m tower plots as well as 26 20m x 20m plots. One elevated trap and one ground trap are deployed for each 400 m^2 plot area.
   3. Ground traps are sampled once per year, while sampling of elevated traps depends on the vegetation that is present (deciduous forest sites are sampled once every 2 weeks during senescence and evergreen sites are sampled once every 1-2 months; some deciduous sites or sites that are hard to access may not be sampled for up to 6 months over the winter).

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
#Dimensions of the Neonics dataset
dim(Neonics)
```

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

6. Using the `summary` function, determine the most common effects that are studied. Why might these effects specifically be of interest?

```r
#Determining the most commonly studied 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 that are studied are population and mortality (the only two effects that are examined by over 1,000 publications). While it would be illuminating to study other effects as well, the health and abundance of insect populations are of ultimate importance; if many members of a species of insect are dying (mortality) or large scale population shifts/changes are occurring, these impacts would most definitively signal the need for action, as they indicate that these insects may not be able to continue providing their current services and benefits.

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.

```r
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 data set are: Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, Italian Honeybee. These insects are all beneficial in certain areas of the world (although a couple of them, the honey bee and the buff tailed bumblebee, have become invasive); the bees are pollinators, while the parasitic wasp helps control populations of insects that harm gardens and 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?

```r
#Finding the class of Conc.1..Author.
class(Neonics$Conc.1..Author.)
```

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

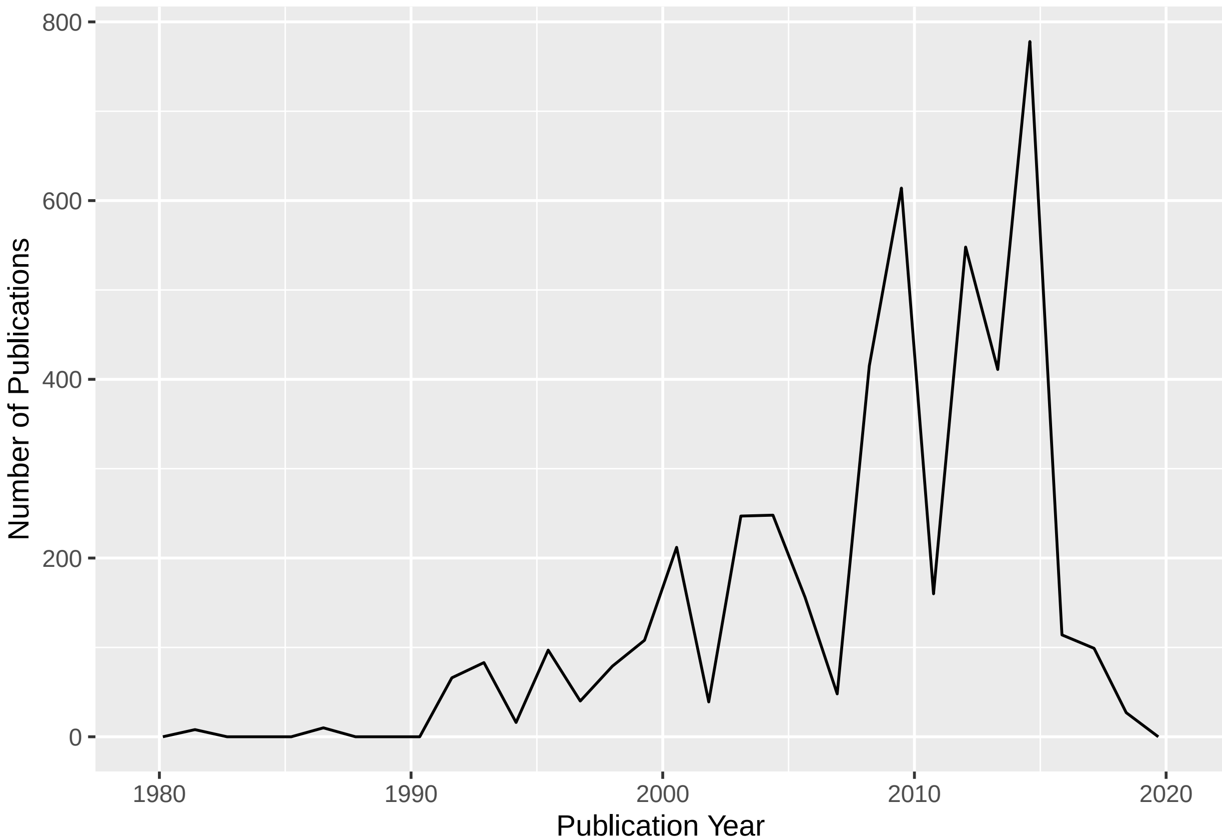
Answer: The class of Conc.1..Author. in the dataset is a factor. Rather than viewing these concentrations as numerical values on which mathematical operations can be performed, this variable is meant to show a limited number of concentration levels that can be seen as categories or groups. In this specific dataset, it is more helpful to see how many studies use specific concentration levels (perhaps for the planning of future studies) as opposed to mathematically manipulating those concentrations.

Explore your data graphically (Neonics)

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

```r
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year)) +
  labs(x = "Publication Year", y = "Number of Publications")
```

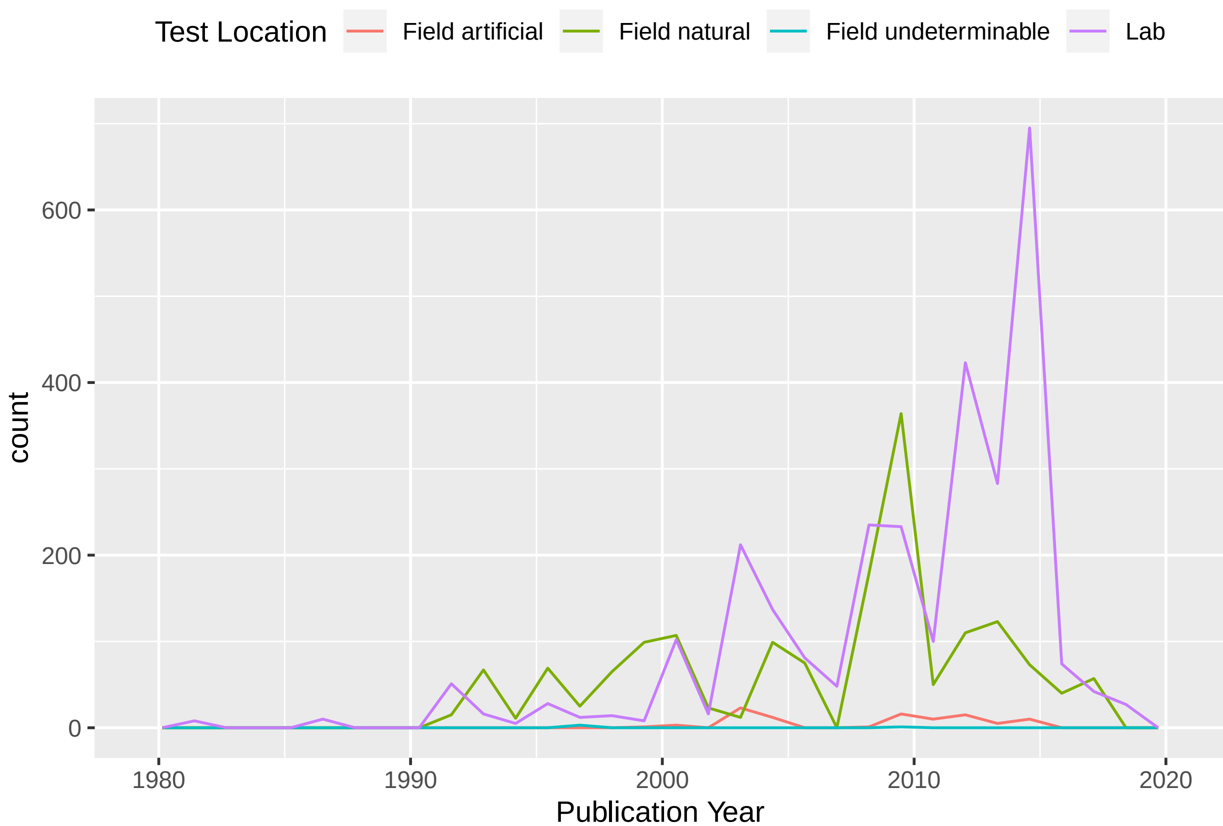
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location)) +
  labs(x = "Publication Year") +
  theme(legend.position = "top") +
  scale_color_discrete(name = "Test Location")
```

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

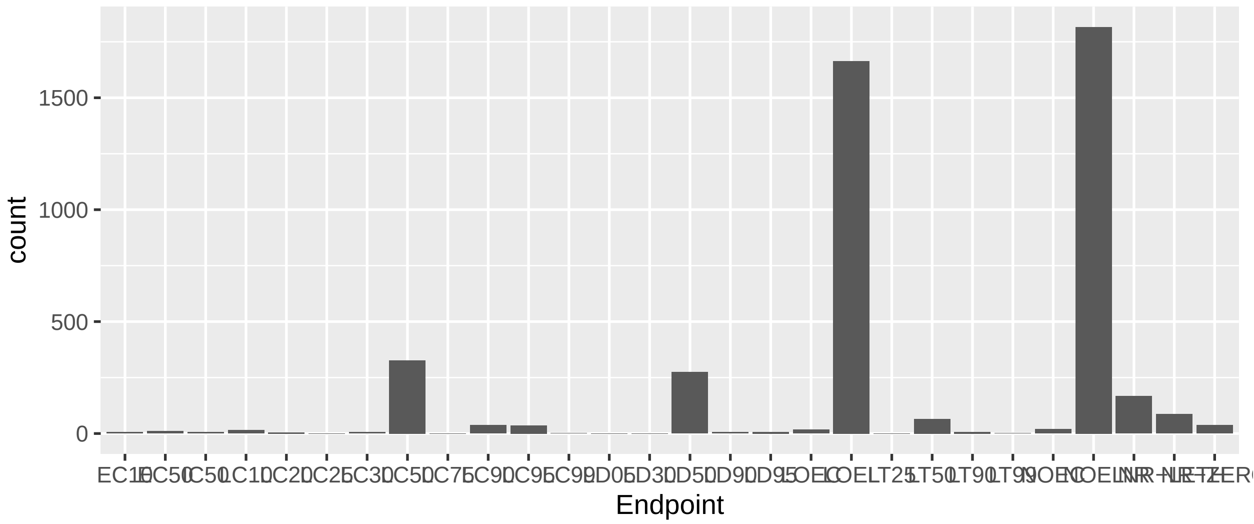


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are the lab and natural field environments. The number of experiments done in the lab matches the overall frequency of publications fairly closely, increasing overall from 1990 to about 2014 (with some decreases also present during this time period), when it abruptly drops. The number of experiments done in the natural field increases and decreases in cycles starting from 1990 to about 2007, spikes in about 2009, and drops off from there.

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()
```



Answer: The most common endpoint is NOEL. NOEL indicates that there is no-observable-effect-level; the highest concentration of the administered chemical produces effects that are not significantly different from the responses of controls (according to the author's reported statistical test). The second-most-common endpoint is LOEL. LOEL represents the category lowest-observable-effect-level; the lowest concentration of the administered chemical produces effects that are significantly different from control responses.

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.

```
#Checking the class of the data
class(Litter$collectDate)
```

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

```
#CollectDate is a factor; changing it to a date:
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")

#Confirming the new class of the variable
class(Litter$collectDate)
```

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

```
#Dates litter was sampled in August of 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`?

```
#Comparing the `unique` and `summary` functions
unique(Litter$plotID)
```

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

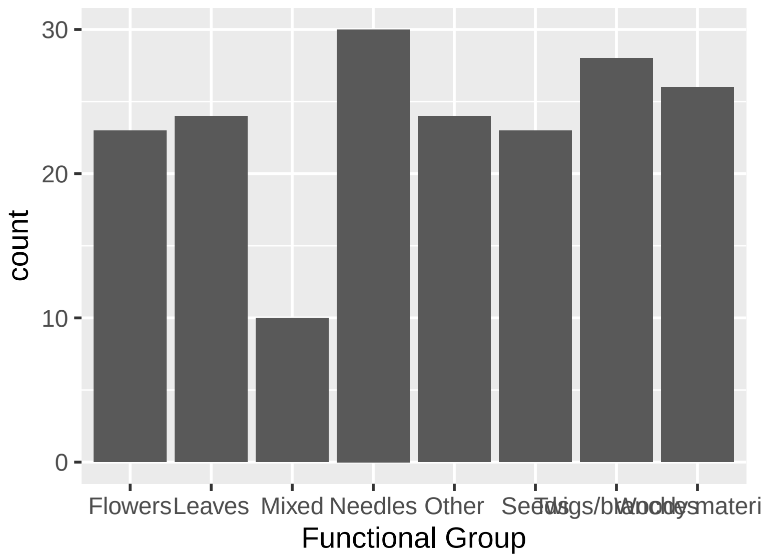
```
summary(Litter$plotID)
```

```
## 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: There were 12 plots sampled at Niwot Ridge. When used on one column of a data frame, `unique` eliminates duplicate values in the column and returns one of each unique value, while `summary` returns each unique value along with its frequency in the column of data.

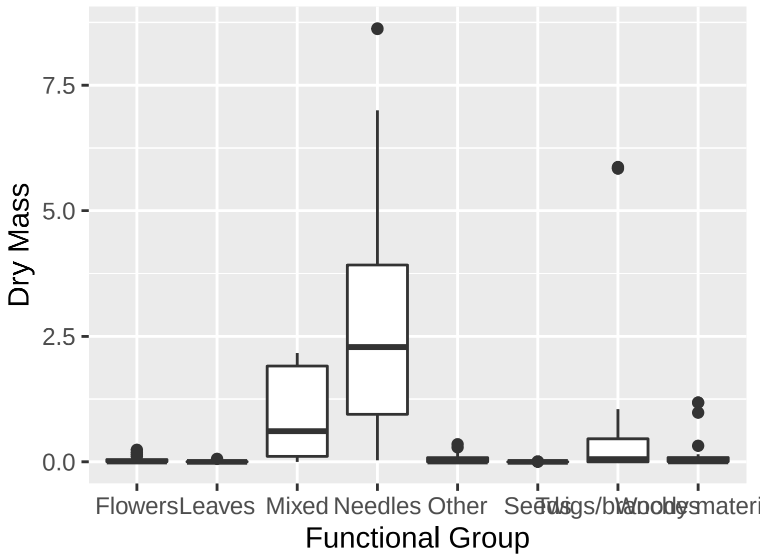
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar() +
  labs(x = "Functional Group")
```



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

```r
#Boxplot
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass)) +
  labs(x = "Functional Group", y = "Dry Mass")
```

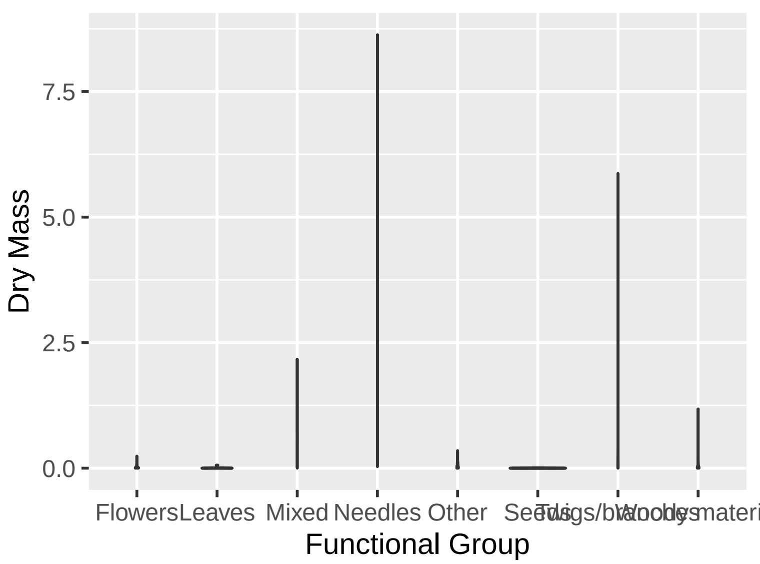



```r
#Violin plot
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass), draw_quantiles = c(0.25, 0.5, 0.75), scale = "coun
  labs(x = "Functional Group", y = "Dry Mass")
```

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

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

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



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

> Answer: Because we are showing the dry mass within each of eight different functional groups, each individual violin plot does not show large enough probability frequencies to make the plots appear as anything much more than a line. The data would need to be divided into fewer groups

for the probability frequency aspect of the violin plot to be informative.

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

> Answer: Needles have the highest biomass, followed by mixed types of litter and twigs/branches (there are some individual samples of woody material that have relatively high dry mass, although woody material as a whole does not).