Assignment 3: Data Exploration

Rachel Landman

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.

```
getwd()
```

```
## [1] "/Users/rmlandman/Desktop/Data Analytics/Environmental_Data_Analytics_2020/Assignments"
```

```
library(tidyverse)

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: Neonicotinoid pesticides are commonly used on plants and are harmful to honey bees and butterflies. Some compare neonicotinoids to DTT because of their role in declining populations of pollinators such as bees and butterflies. It is important to understand the ecotoxicology of neonicotinoids to determine how they harm insects and figure out a sollution that is less toxic. Pollinators are important because they pollinate many of the crops that we rely on for food.

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 are central to forest ecosystems because they provide habitat for organisms and prevent erosion. Additionally, decomposing litter adds valuable nutrients to the soil, which contributes to forest health.

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 woody debris are collected from two traps: elevated and ground. * 1 pair of traps (ground and elevated) is used for every 400m2 plot area. * Ground traps are sampled once per year. Elevated trap sampling frequency depends on type of vegeation at the site.

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, determine the most common effects that are studied. Why might these effects specifically be of interest?

```
summary(Neonics$Effect) ### see all the effects that are studied
```

```
##     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
```

```
summary(Neonics$Effect, maxsum = 4) ### see top 3 effects
```

```
## Population  Mortality   Behavior    (Other)
##       1803       1493        360        967
```

Answer: The top effects that are studied are related to the ability of the species to survive. Population numbers is important to know if a species is endagered or if there are dramatic changes in population from year to year. Mortality is important to know how many insects are dying in order to determine if there is a spike in mortality that could be coorelated with the use of Neonicotinoid pesticides. Behavior is important because it could help determine if the pesticides are negatively impacting the insects without killing them.

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, maxsum = 7)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee      Italian Honeybee
##                   152                   140                   113
##               (Other)
##                  3083
```

```
### said maxsum was 7 because the last value output was "other"
### and I wanted to be able to see the top 6 species in
### addition to the value for "other:
```

Answer: They are all bees, except the one wasp. Bees and wasps are both important for ecosystem functions, as well as food production. Bees are important pollinators and wasps keep pests away.

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: They are factors because some of the rows contain values that are not numeric. For example some rows have / -r ~ and some say NR. This causes R to read that column as categorical data.

Explore your data graphically (Neonics)

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

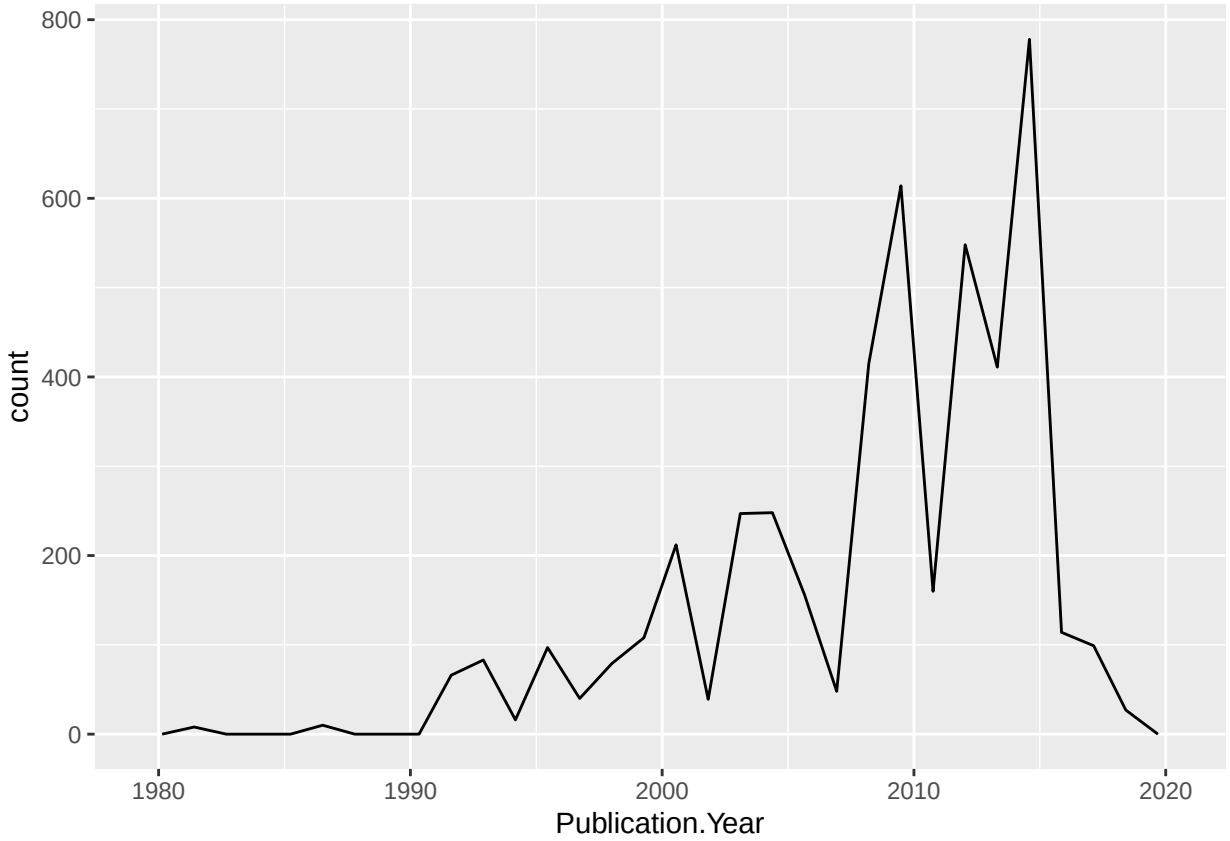
```
colnames(Neonics) ### check to see the column names to ensure using the propper input
```

```
##  [1] "CAS.Number"                        "Chemical.Name"
##  [3] "Chemical.Grade"                    "Chemical.Analysis.Method"
##  [5] "Chemical.Purity"                   "Species.Scientific.Name"
##  [7] "Species.Common.Name"               "Species.Group"
##  [9] "Organism.Lifestage"                "Organism.Age"
## [11] "Organism.Age.Units"                "Exposure.Type"
## [13] "Media.Type"                        "Test.Location"
## [15] "Number.of.Doses"                   "Conc.1.Type..Author."
## [17] "Conc.1..Author."                   "Conc.1.Units..Author."
## [19] "Effect"                            "Effect.Measurement"
## [21] "Endpoint"                          "Response.Site"
## [23] "Observed.Duration..Days."          "Observed.Duration.Units..Days."
## [25] "Author"                            "Reference.Number"
## [27] "Title"                             "Source"
## [29] "Publication.Year"                  "Summary.of.Additional.Parameters"
```

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, bins = 50))
```

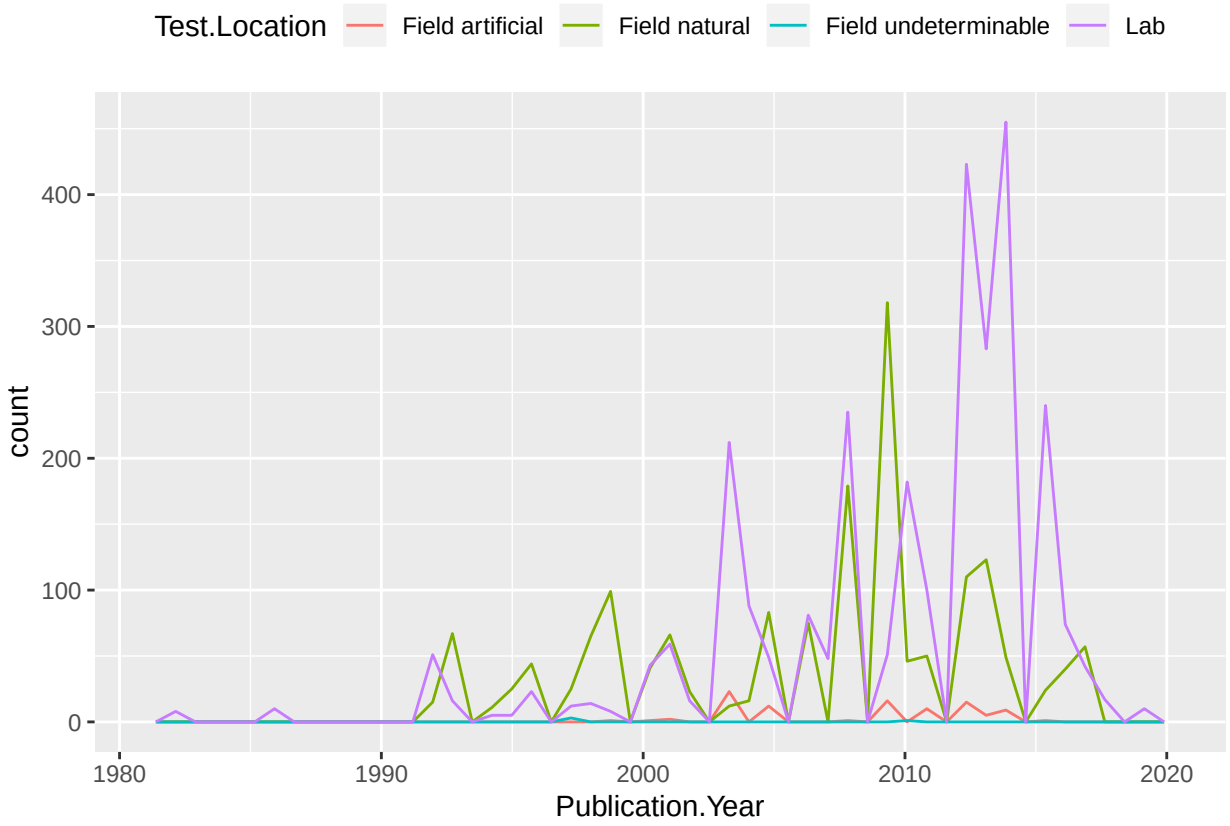
```
## Warning: Ignoring unknown aesthetics: bins
```

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



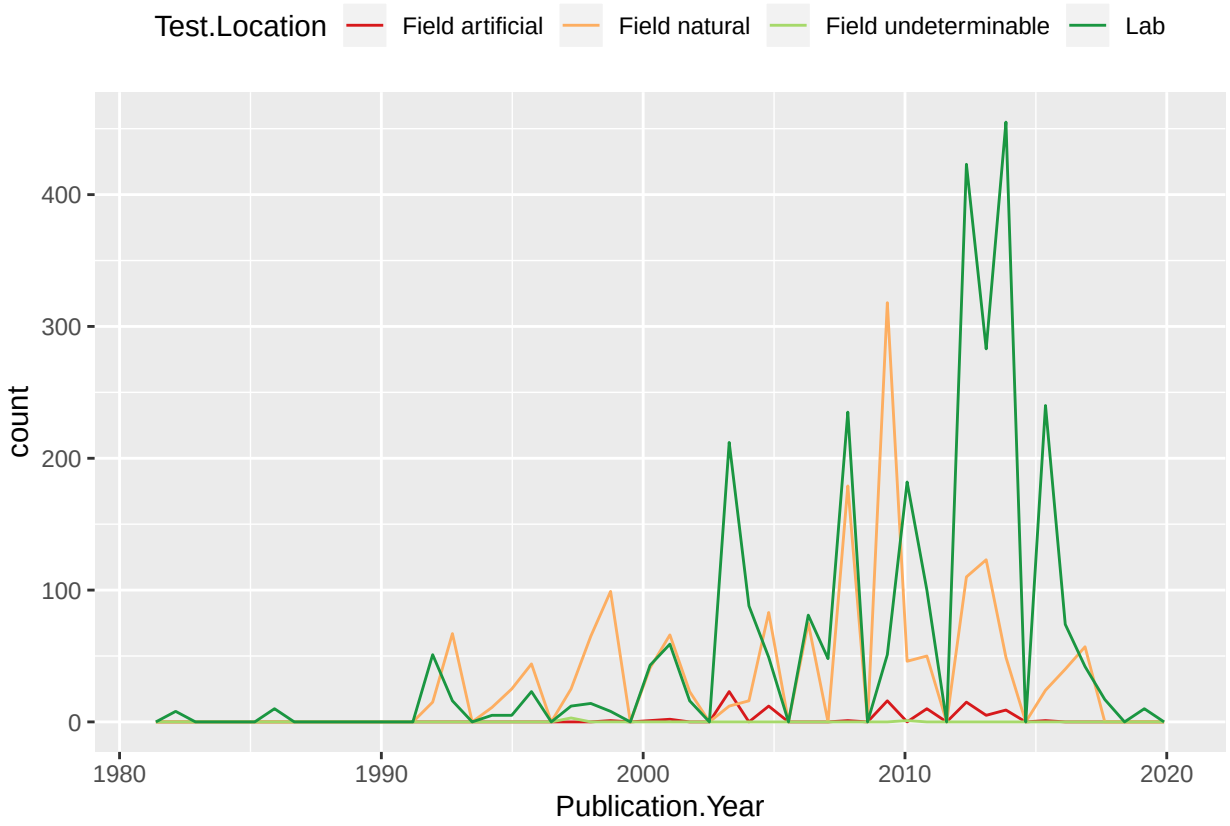
10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 50) +
  theme(legend.position = "top")
```




```
### standard R color scheme

ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 50) +
  scale_color_brewer(palette = "RdYlGn")+
   theme(legend.position = "top")
```



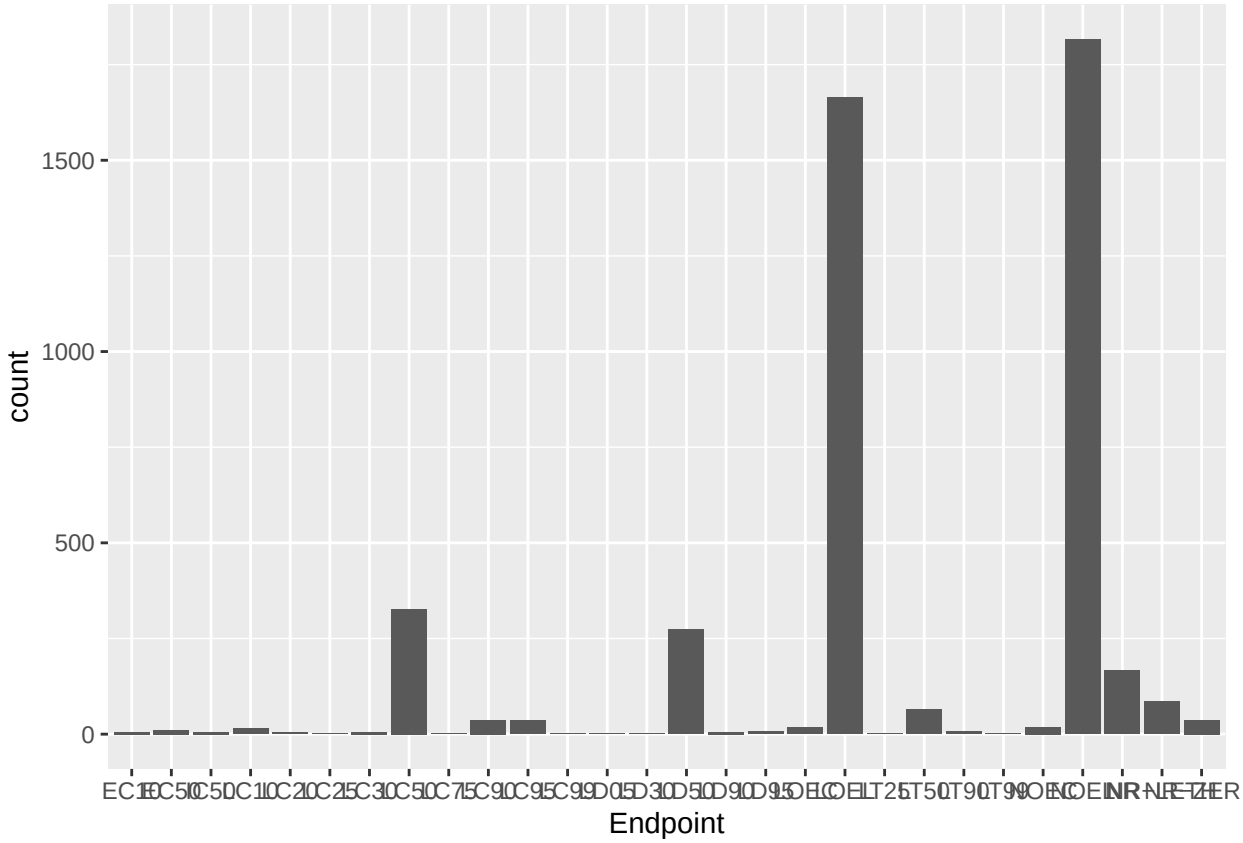

```
### changed color scheme
```

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

> Answer: The lab is pretty consistently the most common test location. Natural field is a bit more common at the beginning and then lab becomes more common. 2009 has a spike in field and drop in lab. Artificial field and undeterminable are very infrequent.

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()
```




```
### couldn't see endpoint names on x axis
### used summary function to find two most common end points

summary(Neonics$Endpoint, maxsum = 3)
```

```
##    NOEL    LOEL (Other)
##    1816    1664    1143
```

Answer: The two most common enpoints are NOEL and LOEL. NOEL is no observable effect level, which means the highest dose did not produce effects that were significantly different than the control. LOEL is lowest observable effect level, which means the lowest dose produced effects that were significantly different than controls.

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, format = "%Y-%m-%d")

class(Litter$collectDate)
```

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

```
unique(Litter$collectDate, incomparables = 2018-8)
```

```
## [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`?

```
colnames(Litter)
```

```
##  [1] "uid"                    "namedLocation"
##  [3] "domainID"               "siteID"
##  [5] "plotID"                 "trapID"
##  [7] "weighDate"              "setDate"
##  [9] "collectDate"            "ovenStartDate"
## [11] "ovenEndDate"            "fieldSampleID"
## [13] "massSampleID"           "samplingProtocolVersion"
## [15] "functionalGroup"        "dryMass"
## [17] "qaDryMass"              "remarks"
## [19] "measuredBy"
```

```
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
```

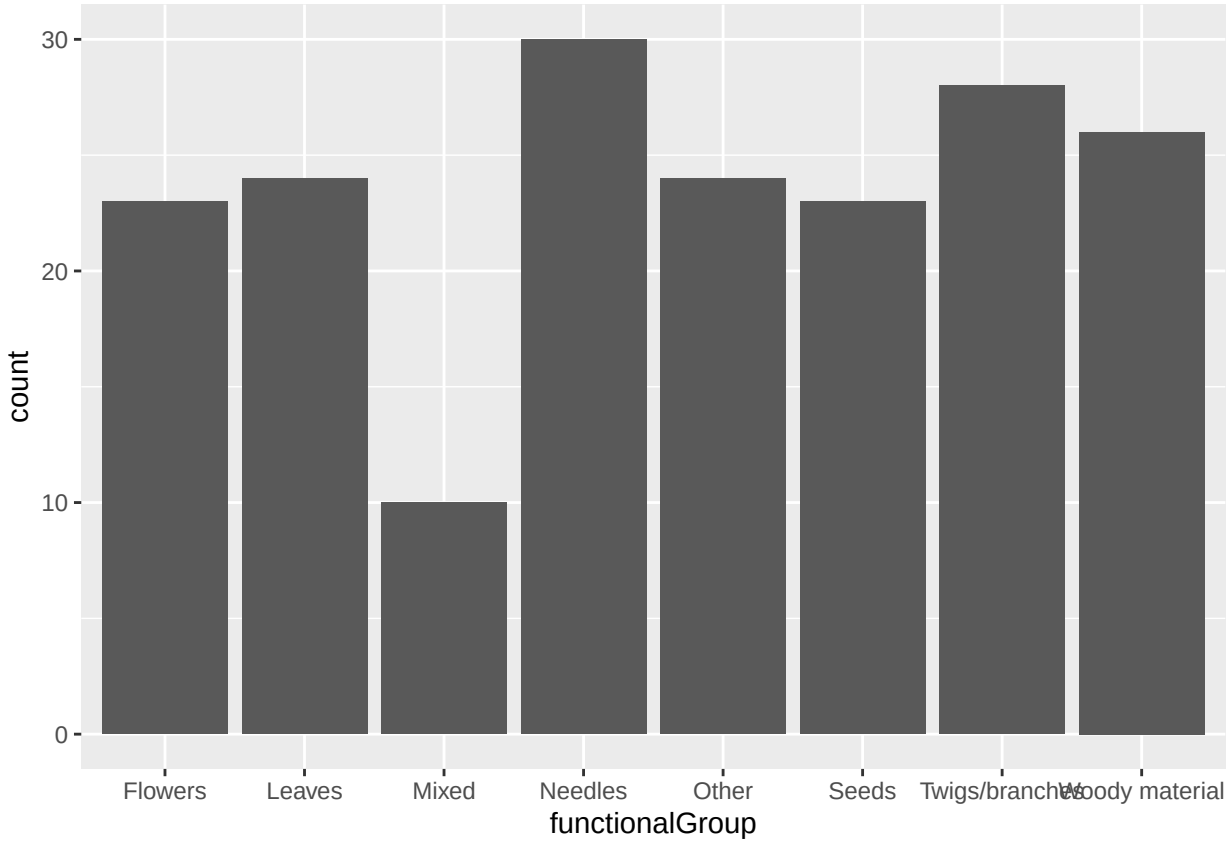
```
summary(Litter$plotID)
```

```
## 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: Unique tells you how many plots were sampled and summary tells you how many plots were samples and how many samples were taken at each plot

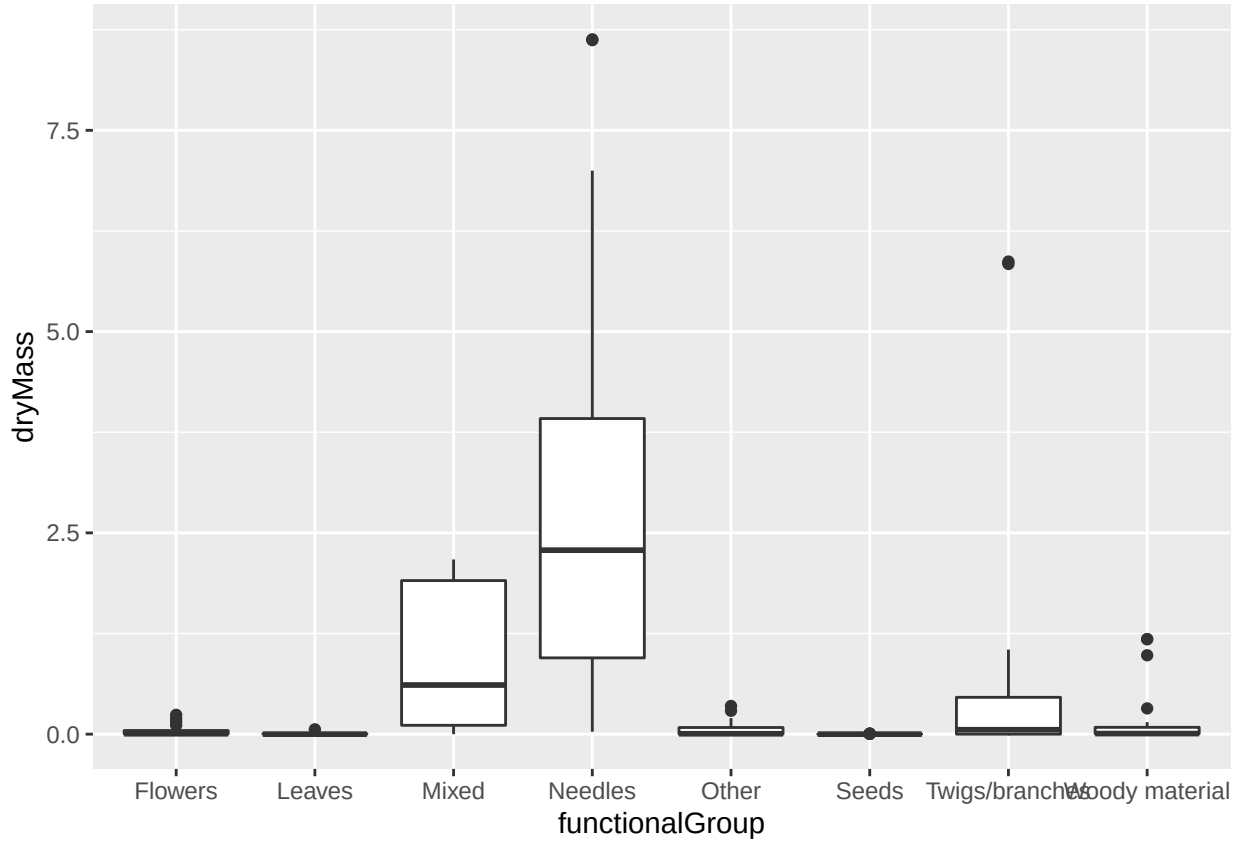
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
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.

```
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass))
```

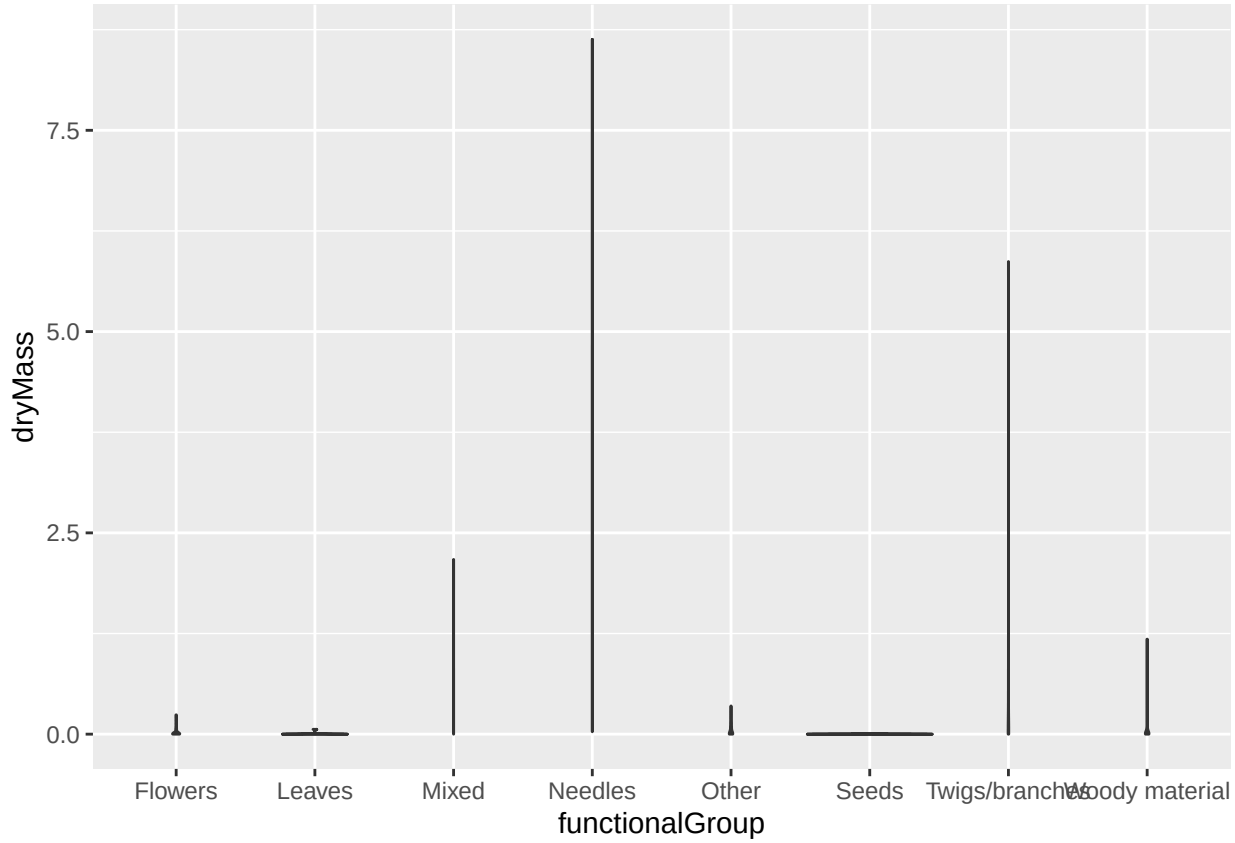



```
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75))
```

```
## 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: It is much easier to see the median and the interquartile range in the boxplot. Because the dry mass numbers are all so different, the violin does not get wide, but rather looks like a straight line. Just seeing the data as a straight line make it difficult to determine where the majority of the points are within the given range. Therefore it is hard to compare the different functional groups. We also saw that there was a rather even distriubtion among each group when we made the bar graph and therefore don't need the violin plot which is better at expressing difference in count.

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

> Answer: Needles and Mixed litter tend to have the hights biomass at these sites.