Assignment 3: Data Exploration

Monisha Eadala

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.

```
# To set up working directory
getwd()

## [1] "/Users/monishaeadala/Environmental_Data_Analytics_2020"

# To load packages
library(tidyverse)

# To import 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: We are interested in the ecotoxicologoy of neonicotinoids on insects since we want to know the effects of neonicotinoids on both target organisms and non-target organisms. From glancing at the Neonics dataset, we can tell that neonicotinoids seem to be effecting some of the beneficial insects and the harmful pests in a similar fashion (by causing mortality). For example, even though they seem to be causing mortality to the target pests such as coffee bean weevils and sweetpotato whiteflies, they are also causing the same harm to beneficial insects such as honey bees, bumble bees and few of the lady beetle species. Also it is necessary to have and understand such data that could aid in controlling the dossage and concentration of the pesticide according to the level of pest control required without going overboard causing harm to the ecological balance.

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: The litter and woody debris data may be used to estimate the annual Aboveground Net Primary Productivity (ANPP) and aboveground biomass at plot, site, and continental scales that are important to understand vegetative carbon fluxes over the time and answer very important questions regarding the global carbon balance, the location of the missing carbon sink, and predictions of global climate change. Such knowledge and predictions can influence current management practices and public policy.

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 are collected individually from two different traps - elevated and ground traps. All masses reported after processing are reported at the spatial resolution of a single trap and the temporal resolution of a single collection event. * Spatial Sampling: For elevated traps, litter and fine woody debris sampling is executed at terrestrial NEON sites that contain woody vegetation more than 2m tall. Locations of tower plots are selected randomly within the 90% flux footprint of the primary and secondary airsheds. * In sites with forested tower airsheds, the litter sampling is targeted to take place in 20 40m x 40m plots. In sites with low-statured vegetation over the tower airsheds,litter sampling is targeted to take place in 4 40m x 40m tower plots (to accommodate co-located soil sampling) plus 26 20m x 20m plots. Also, Trap placement within plots may be either targeted or randomized, depending on the vegetation. In sites with more than 50% aerial cover of woody vegetation more than 2m in height, placement of litter traps is random. * Temporal Sampling: Ground traps are sampled once per year. Target sampling frequency for elevated traps varies by vegetation present at the site, with frequent sampling (1x every 2weeks) in deciduous forest sites during senescence, and infrequent year-round sampling (1x every 1-2 months) at evergreen sites.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) # Gives the dimensions (number of rows and columns) of the dataset (in this the number of rows are 4623, while the number of columns are 30)
```

```
## [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) # Lists all the effects with their respective
frequency; the ones with the most frequency are the most common.
```

```
##     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”, “mortality”, “behavior”, “feeding behavior”, “reproduction”, and “development”. These effects might specifically be of interest to understand what are some of the most common effects of the insecticides and to even determine if such effects are detrimental to the environment. Such understanding can form a basis for environmental decision analysis and can determine the future use of such insecticides.

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) # Lists the studied species with the
most common ones listed at the top
```

```
##                          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 studies species in the dataset are “honey bee”, “parasitic wasp”, “buff tailed bumblebee”, “Carniolan honey bee”, and “bumble bee” and “Italian honeybee”. the one thing all these species have in common is that they are all considered beneficial insects that feed on other pests that harm the crops. They all have good reputation in the field of garden and crop management/ pest control.

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.) # Generates the class of Conc.1..Author.
```

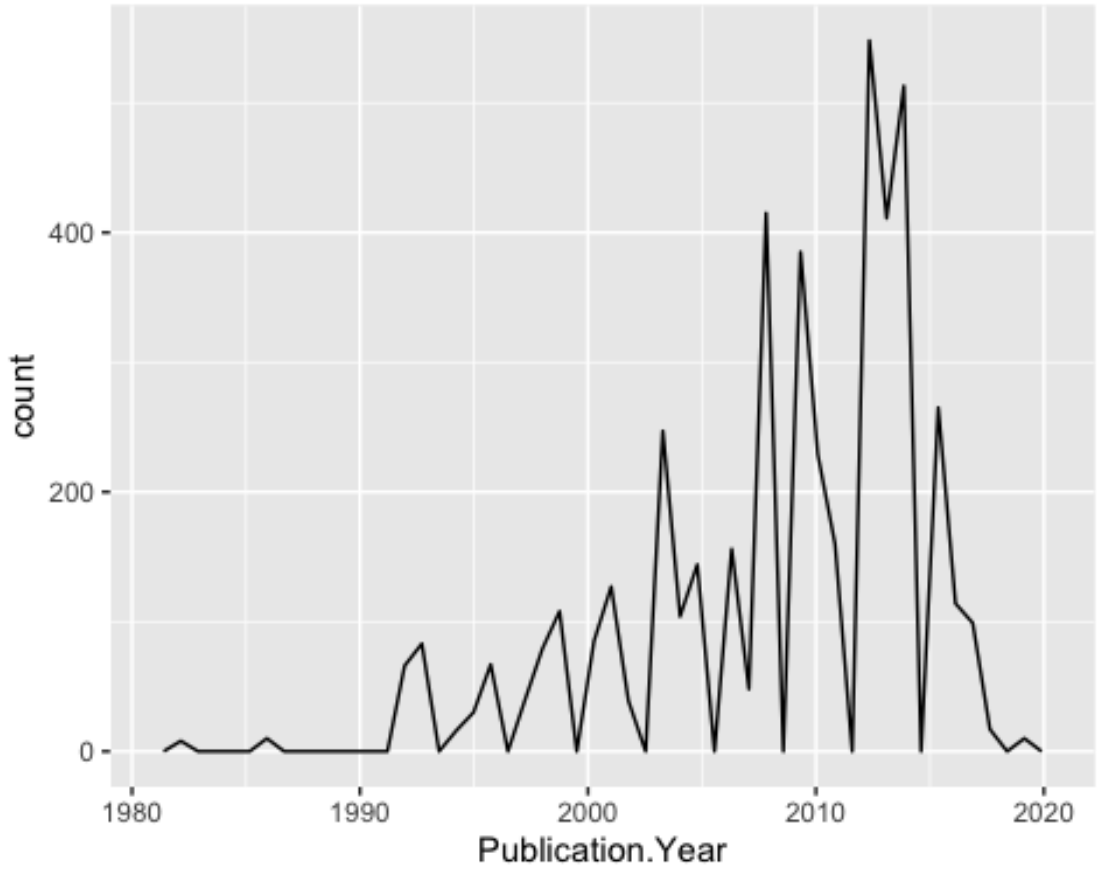
```
## [1] "factor"
```

Answer: The class of Conc.1..Author. is “factor” and not numeric, and this is because the values cannot be considered numeric yet given that they all have different units. Only after converting all the units to a single consistent unit, we should convert the class of the variable to numeric.

Explore your data graphically (Neonics)

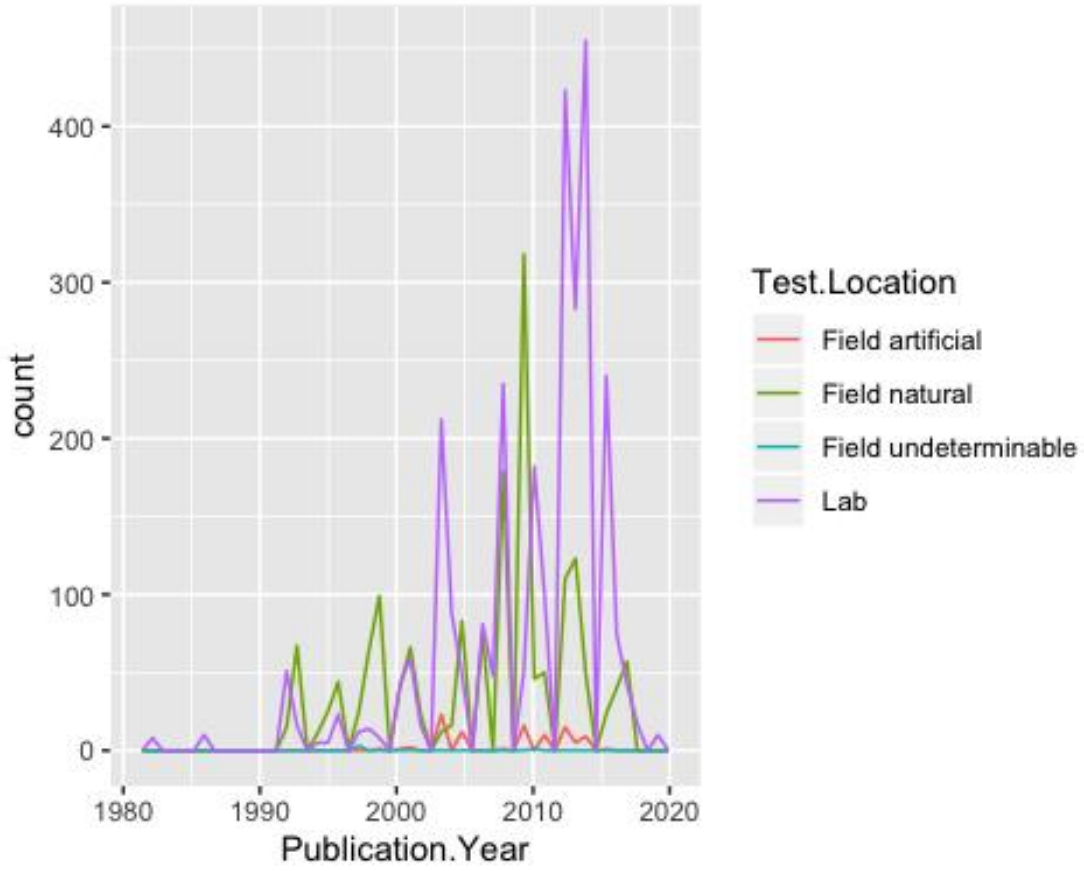
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50) # Generates a plot of
the 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.

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 50)
# Reproduces the same graph but now add a color aesthetic so that different
Test.Location are displayed as different colors
```

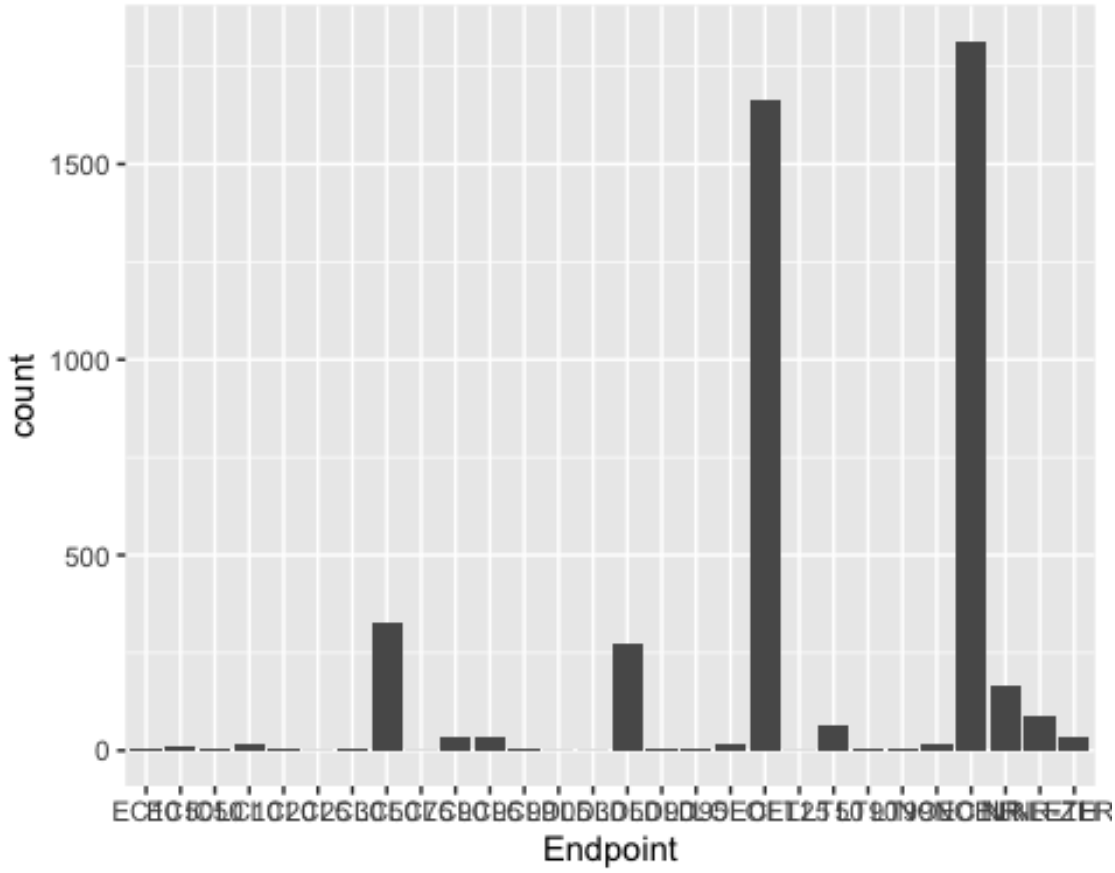


Interpret this graph. What are the most common test locations, and do they differ over time?

Answer: The most common test locations overall according to the graph seem to be Lab, followed by Field natural, Field artifical and then Field underdeterminable. However, they do differ over the time. For instance, right before 2010 (around 2009) Field natural seems to be more common, while right after 2010 Lab seems to be more common than Field natural.

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() # Creates a bar graph of Endpoint counts
```



Answer: The two most common endpoints are NOEL (most common) and LOEL (second-most common). NOEL is the abbreviation for No-observable-effect-level, and it is defined as the highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test (NOEAL/NOEC). LOEL is the abbreviation for Lowest-observable-effect-level, and it is definied as the lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC).

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) # Determines the class of collectDate (in this case it is a "factor")

## [1] "factor"

Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d") # Changes it to a date
class(Litter$collectDate) # Confirmed that the new class of collectDate is a "Date"

## [1] "Date"
```

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?

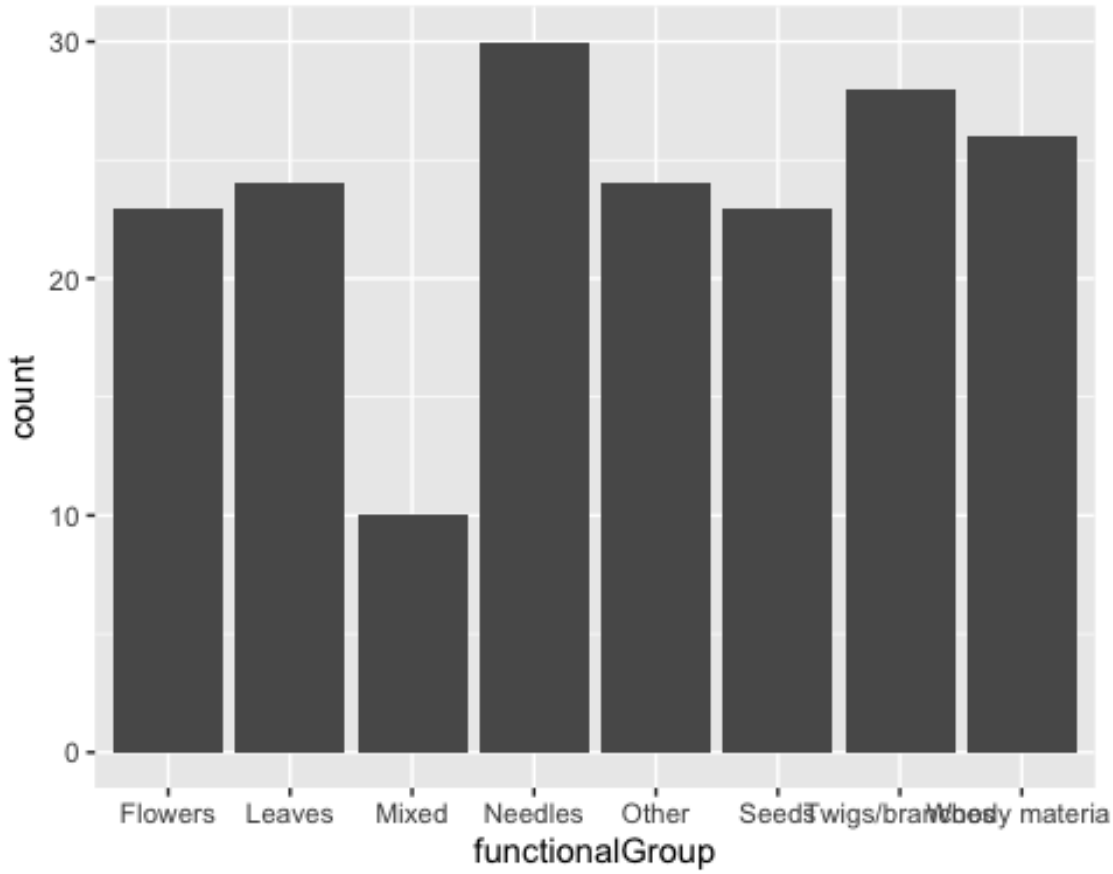
```
unique(Litter$plotID) # Determines how many plots were sampled at Niwot Ridge without duplication

##  [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
```

Answer: The 'unique' function eliminates duplicate elements/rows from a vector, data frame or array, while the 'summary' function does not. In other words, the 'unique' function strictly returns the unique rows in a dataframe while the summary returns all the rows in a dataframe. For the 'unique' function, all the elements in each row must match in order to be termed as duplicates.

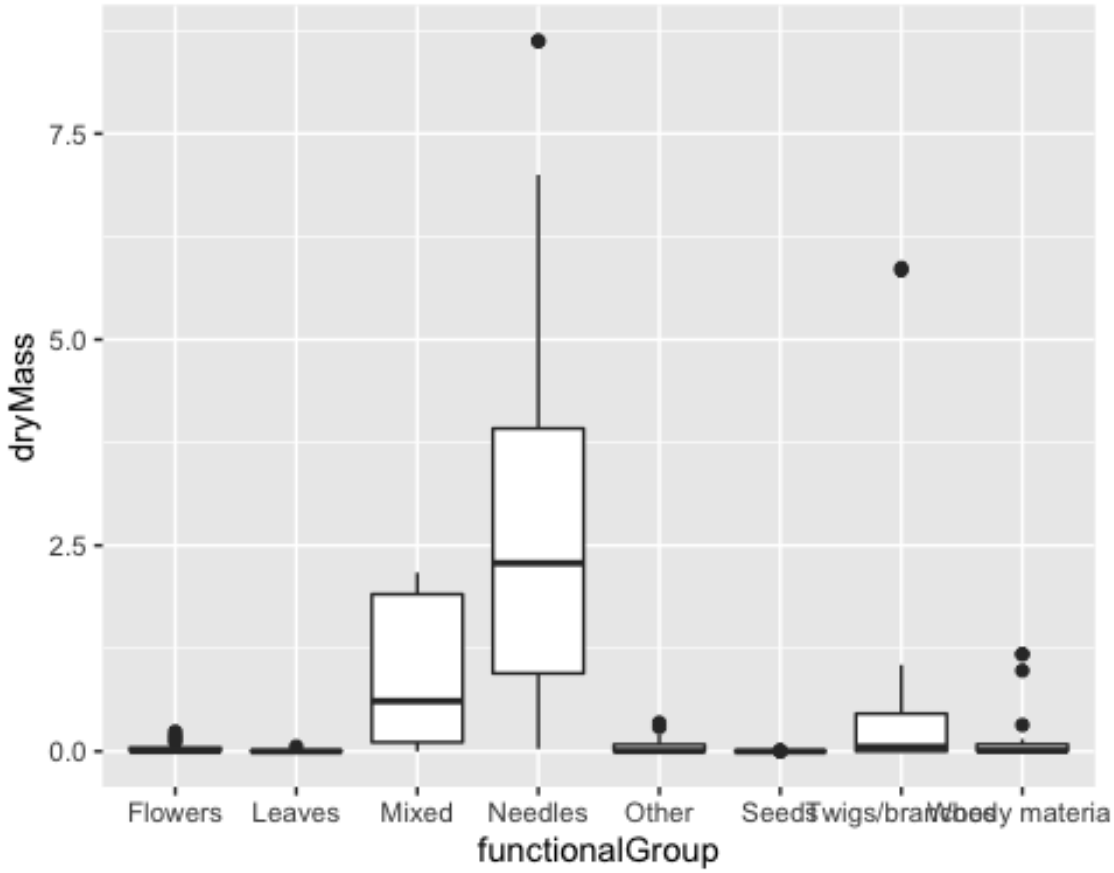
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar() # Creates a bar graph of functionalGroup counts
```



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)) # Creates a boxplot of
dryMass by functionalGroup
```

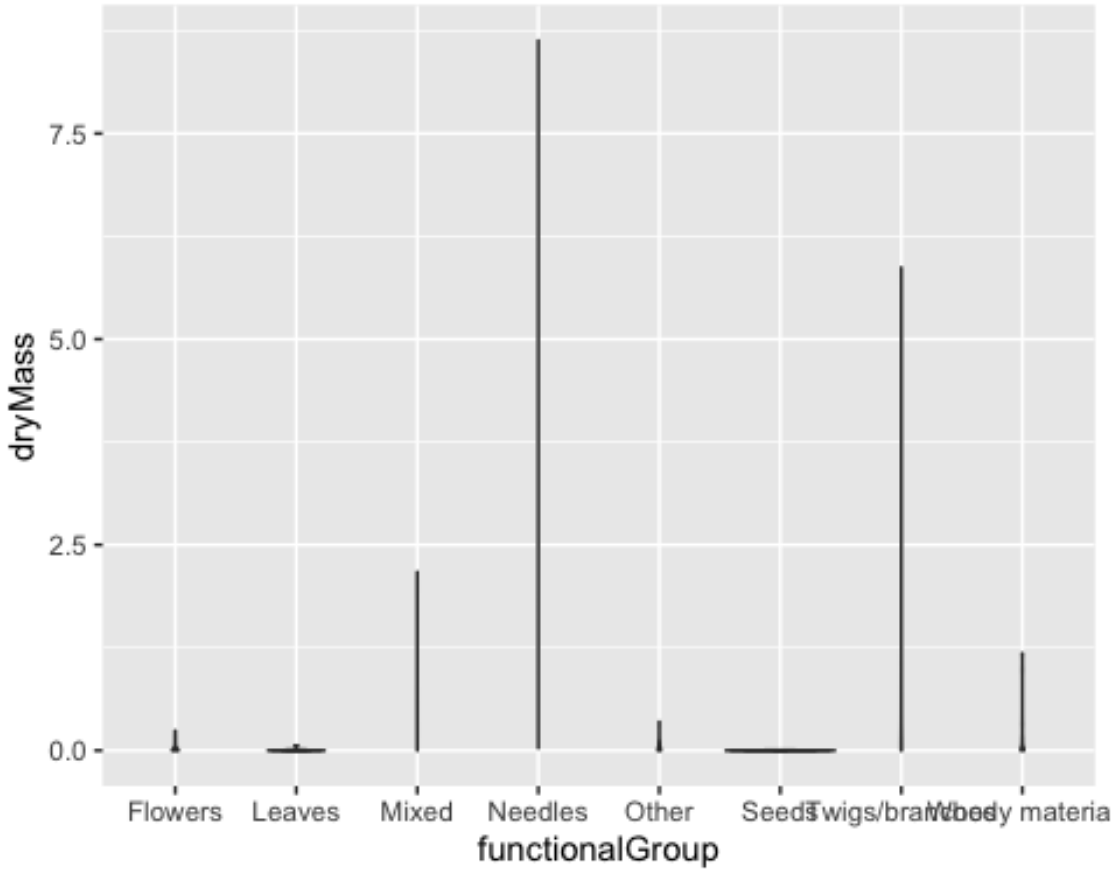



```
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75)) # Creates a violin plot of
dryMass by functionalGroup

## 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: A boxplot is a more effective visualization option than the violin plot in this case because the boxplot is giving us more information; especially regarding the median, first quartile, third quartile and outliers. On the other hand, the only information the violin plot is giving us is the minimum and the maximum observations (which the box plot gives us as well). Also, R tells us that there are 3 counts of collapsing due to unique 'x' values in the violin plot, which makes the visual even more unappealing. Therefore, a boxplot is much better in this case.

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

Answer: From the box plot it is evident that Needles tend to have the highest biomass at these sites. After Needles, Mixed litter seems to have the next highest biomass.