Assignment 3: Data Exploration

Sierra Kindley

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() #check working directory
```

```
## [1] "C:/Users/sierr/Documents/Duke University/ENVIRON-872L/Environmental_Data_Analytics_2020"
```

```
library(tidyverse) #load tidyverse
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv") #upload ecotox
#neonicotinoid dataset and name it 'Neonics'
Litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv") #upload Niwot
#Ridge NEON dataset and name it '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 ecotoxicologoy of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: It is important to evaluate the ecotoxicology of neonicotinoids on insects in order to evaluate whether said insecticides are effective in serving their purpose (killing insects in agricultural environments to protect crops) and, if so, their mechanism(s) of action (how they kill insects). In addition, this information might be useful to help evaluate the overall toxicity of neonicotinoids in the environment or in terms of how they may effect humans or other species if they were to come in contact with said substances. Is it really safe to use these on or around crops that may be consumed?

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: Aside from serving as a nesting material for some species, decomposing leaf litter and woody debris release nutrients into the soil and keep it moist. Therefore, they are important components of healthy soil. Studying litter and woody debris that falls to the ground in forests can aid in our understanding of the underlying soil and help us draw conclusions about its overall 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 fine woody debris are collected from elevated traps and ground traps, respectively.* Ground traps are sampled once every year. Sampling frequency for elevated traps is dependent on vegetation type present at the site. Sampling occurs once every two weeks in deciduous forest sites and once every one to two months at evergreen sites. *All masses of litter and fine woody debris reported after processing of the material are reported at the spatial resolution of a single trap and the temporal resolution of a single collection event.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) #ask R what the dimensions of the dataset 'Neonics' are
```

```
## [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) #ask R for a summary of the 'Effect' column in the dataset 'Neonics'
```

```
##     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 studied are population and mortality. It is important to know how/if neonicotinoids affect insect populations and/or kill individual insects. If the aim at using neonicotinoids in agriculture is to kill or deter pests, studying mortality and populations of insects will aid in determining how effective their use is. In addition, these types of studies may shed light on potential undesired effects of neonicotinoid use on insects (killing or negatively impacting populations of essential insect species).

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)
```

```
##                          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
```

```
#ask R for a summary of the 'Species.Common.Name' column in the dataset 'Neonics'
```

Answer: The six most commonly studied species in the dataset include the Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. All of these species benefit agriculture (and therefore humans) in some way. The honey bees and bumblebees are essential pollinators, and the Parasitic Wasp naturally controls agricultural pests.

As these species are vital to the production of healthy agriculture, they are studied more frequently. It is important to understand the effects of neonicotinoids on these species, as harming them is not at all what we desire to do (yet likely, given the toxicity and environmental persistence of neonicotinoids).

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"
```

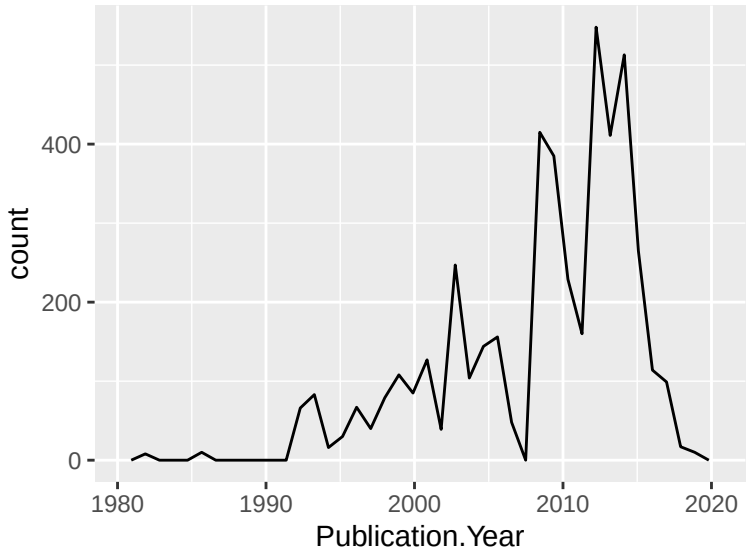
```
#ask R the class of the column 'Conc.1..Author.' in the dataset 'Neonics'
```

Answer: The class of 'Conc.1..Author.' is factor. 'Conc.1..Author.' is not numeric, as all of the values in the column are not numeric. 'Conc.1..Author.' contains numeric values as well as some character values. Therefore, it cannot be classified as numeric and must be classified as factor instead.

Explore your data graphically (Neonics)

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

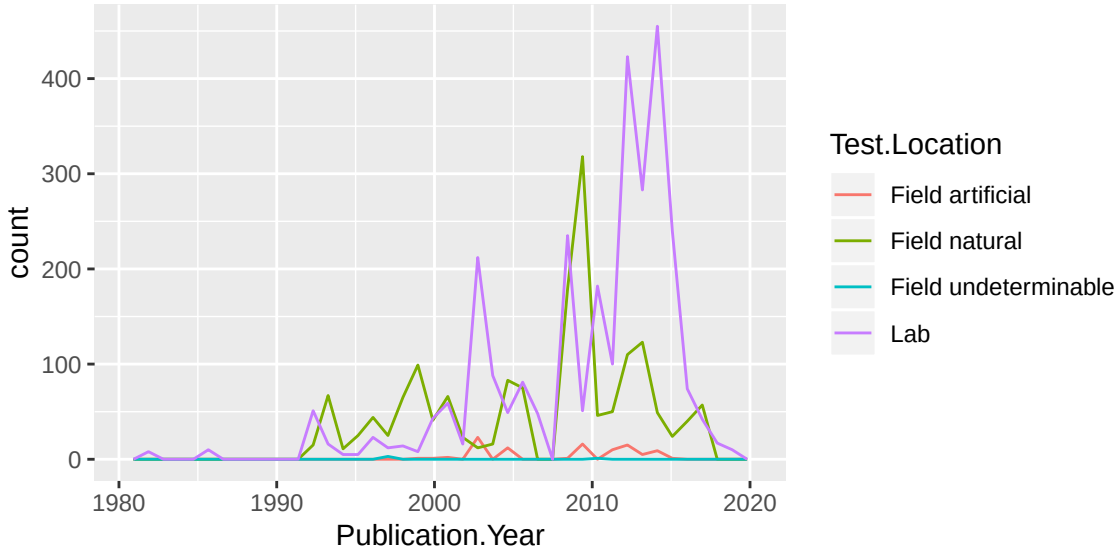
```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 40)
```




```
#generate frequency line graph with 40 bins 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 = 40)
```



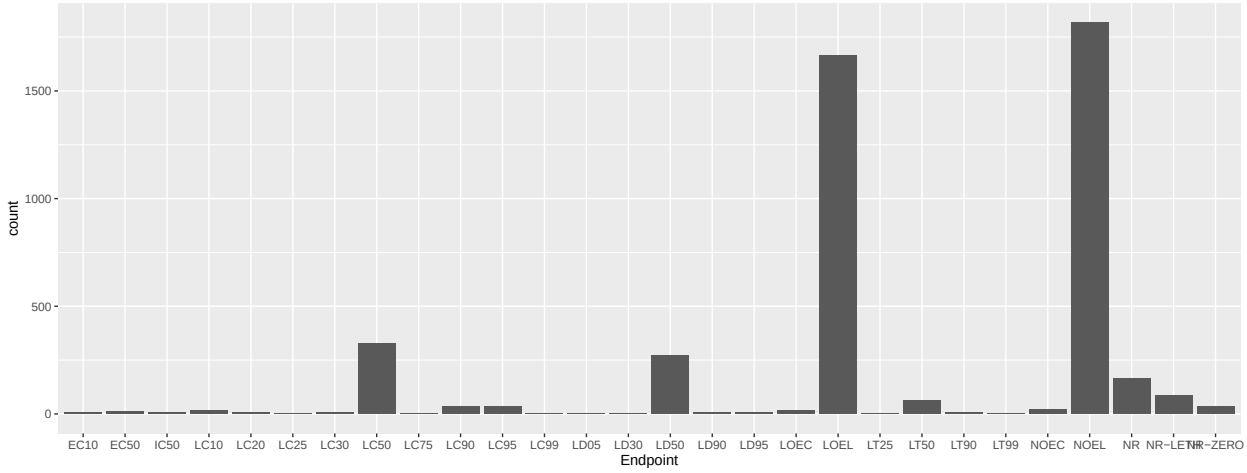

```
#generate frequency line graph with 40 bins of the number of studies conducted by publication year;
#different test locations are displayed as different colors
```

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

> Answer: The most common test locations are 'lab' and 'field natural'. These two test locations are consistently the two most common of the four accounted for. However, over time, they altnerate being the single most common test location. The pattern for the most common test location generally goes field natural, lab, field natural, lab, with a very large spike in lab tests and a relatively significcant drop in field natural tests from approximately 2010 to 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.

```
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar()
```




```
#generate bar graph of Endpoint counts
```

> Answer: The two most common end points are LOEL and NOEL. LOEL, or lowest observable effect level, is defined as the lowest dose (concentration) of a substance producing effects that are significantly different (as reported by authors) from responses of controls. NOEL, or no observable

effect leve, is defined as the highest dose (concentration) of a substance producing effects that are not significantly different from respoonses of controls (according to author's 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.

```
class(Litter$collectDate) #ask R what the class of collectDate is
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
#change the class of collectDate to date with the format "%Y-%m-%d"
class(Litter$collectDate) #ask R what the class of collectDate is again
```

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

```
unique(Litter$collectDate, Litter$collectDate >= "2018-08-01"
       & Litter$collectDate <= "2018-08-31", incomparables = FALSE)
```

```
## [1] "2018-08-02" "2018-08-30"
```

```
#extract all the dates on which litter was sampled in August 2018 from the dataset with no duplicates
```

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, incomparables = FALSE)
```

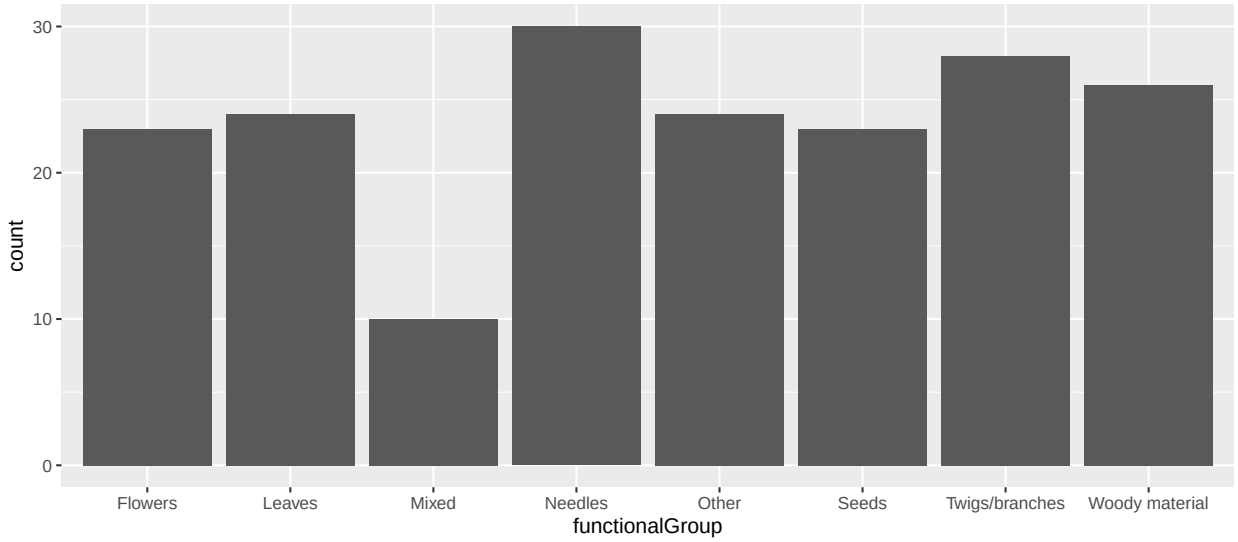
```
##  [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

```
#ask R for only the names of the plots sampled at Niwot Ridge with no duplicates
```

Answer: Twelve (12) plots were sampled at Niwot Ridge. The 'unique' function generates a list of each of the plots sampled at Niwot Ridge (their names only, with no duplicates listed), while the 'summary' function produces a list of each of the plots sampled at Niwot Ridge along with a number denoting how many times they appear in the dataset.

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.

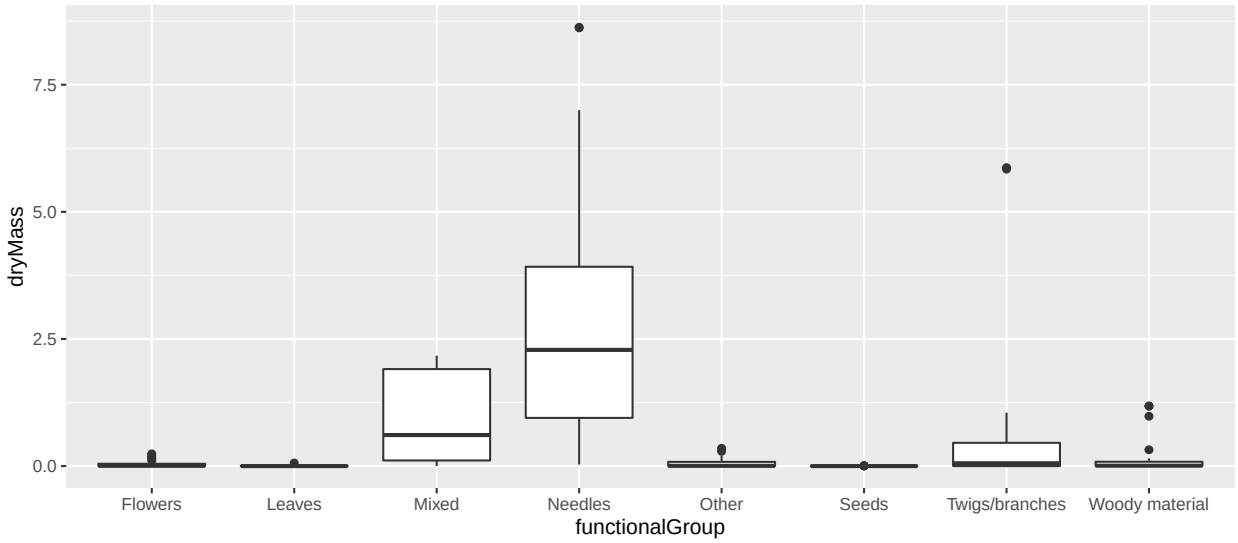
```
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar()
```




```
#generate a bar graph of functionalGroup counts
```

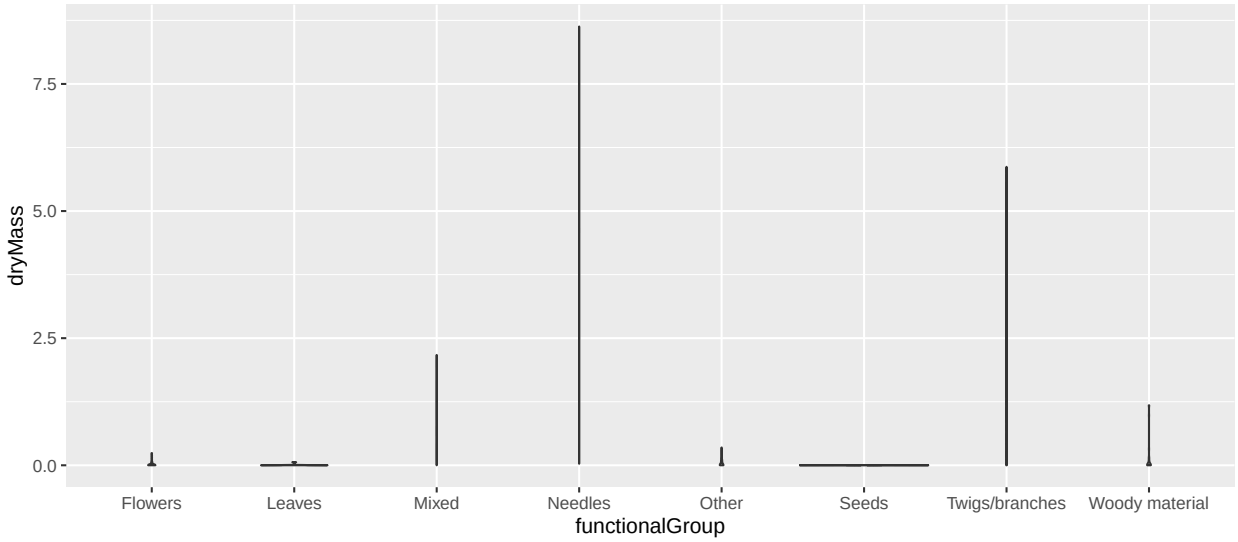
15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




```
#generate a boxplot of dryMass by functionalGroup

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




```
#generate a violin plot of dryMass by functionalGroup
```

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

> Answer: In this case, the spread/variability of the data within each functionalGroup is relatively low. Thus, a boxplot is a more effective visualization option than the violin plot, as it actually allows us to see the data and its corresponding quartiles displayed on the graph. The violin plot does not allow us to visually view much, and the data is very difficult to see (simply looks like a single line and a point or two for each functionalGroup).

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

> Answer: Needles tend to have the highest biomass at these sites. This is then followed by mixed litter and twigs/branches.