Assignment 3: Data Exploration

Danlei Zou

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A03_DataExploration.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change "Student Name" on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.
6. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai.

The completed exercise is due on Sept 30th.

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 include the subcommand to read strings in as factors.

```
getwd()
```

```
## [1] "/Users/danleizou/EDA-Fall2022"
```

```
## [1] "/Users/danleizou/EDA-Fall2022"

#creating datasets
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = TRUE)
Litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv", stringsAsFactors = TRUE)
```

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: We might be interested to see how certain species of insects react to neonicotinoids. This data would help keep track of how effective the insecticides are against different species. Some may be killed by the usage of such insecticides, while others may end up adapting to live amidst the usage or even become immune to them.

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: Wood litter and debris that fall to the forest ground add nutrients to the soil, and provide cover for other animals that live in the forest. The amount of wood litter and debris may also be indicative of the overall health of the forest and the trees in it.

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 fine woody debris samplign is executed at terrestrial NEON sites that contain woody vegetation over 2m tall. 2. Litter and fine woody debris are collected from elevated and ground traps, respectively. 3. In sites with > 50% aerial cover of woody vegetation >2m in height, placement of litter traps is random.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#checking dimensions of Neonics dataset
dim(Neonics)
```

```
## [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?

```
#checking summary of Effect column in dataset
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: Population and Mortality are the most common effects studied. These might be of interest in order to keep track of how the population of each species changes over time as the insecticide is introduced and throughout its future usage.

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.

```
#checking most commonly studied species
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 common species studied are the Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumbe Bee, and Italian Honeybee. These are all pollinator insects, with 5 of them being bee species.

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
#checking Conc.1..Author class
class(Neonics$Conc.1..Author.)
```

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

```r
## ["factor"]
```

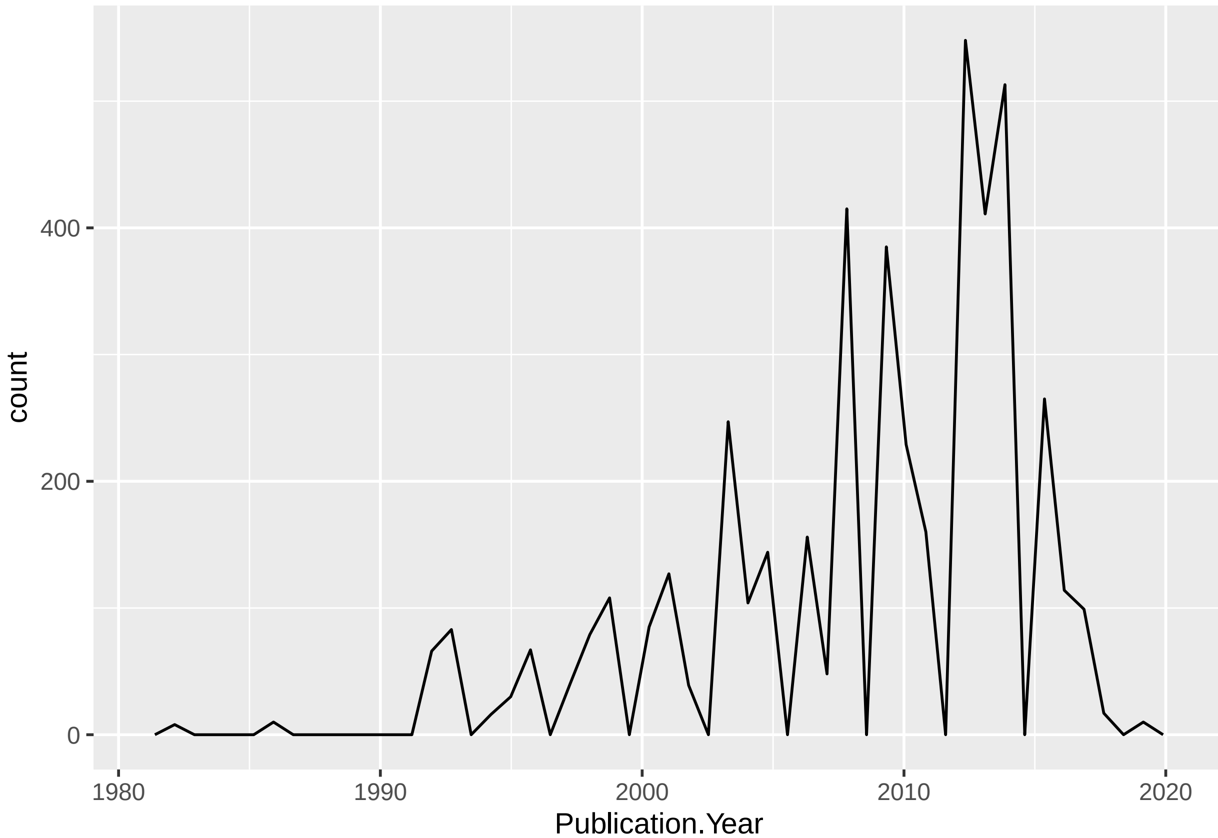
> Answer: The class of Conc.1..Author. is a character class because the data is categorical, not numeric.

Explore your data graphically (Neonics)

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

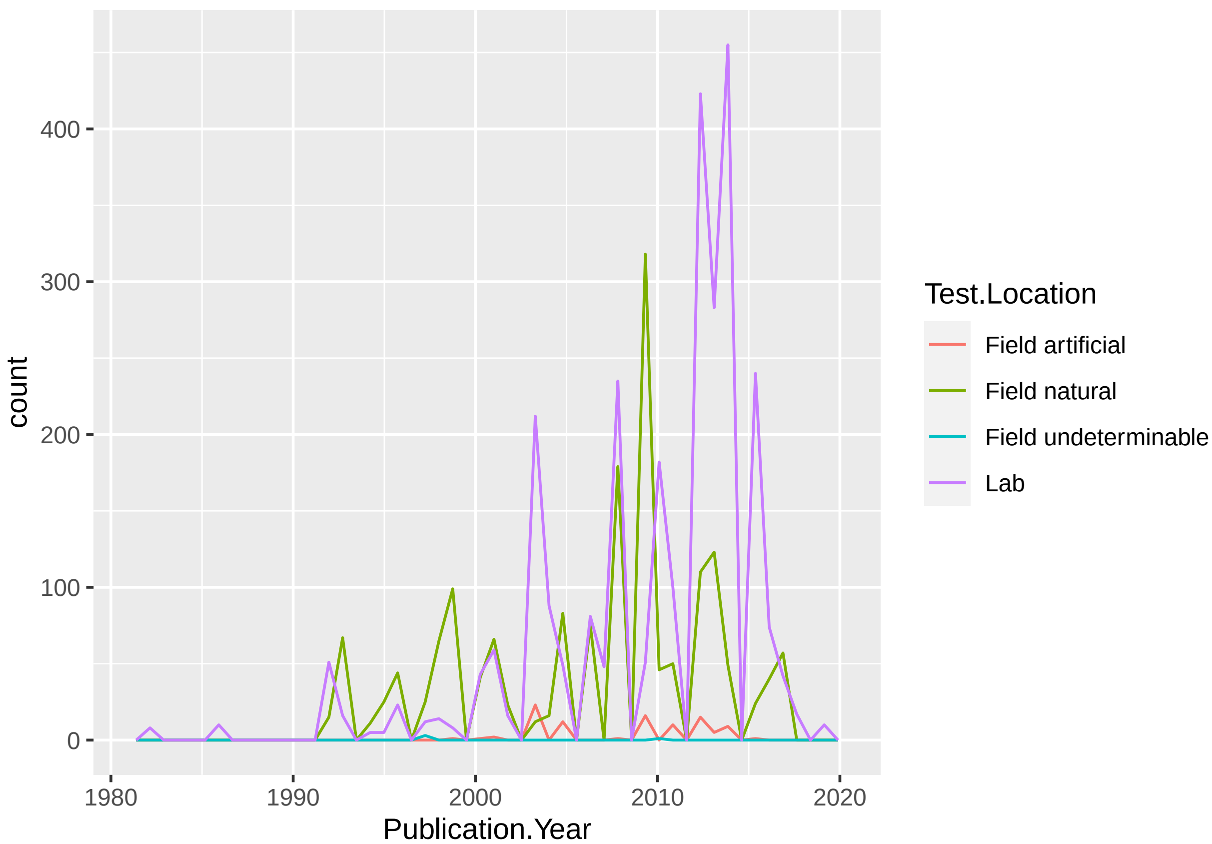
```r
#creating line plot by pub year
library(ggplot2)

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



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

```
#adding color to line plot by pub year
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 50)
```

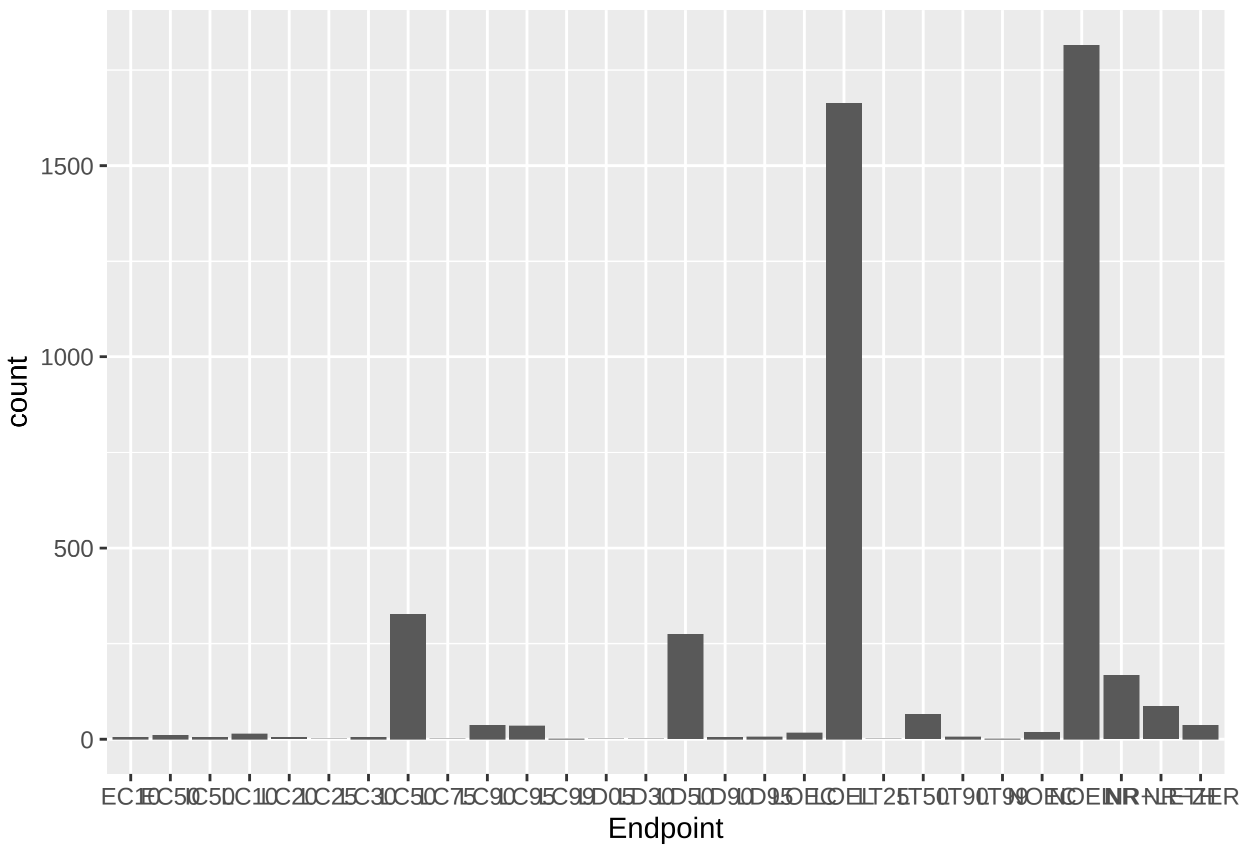


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are labs, peaking at its highest around 2013/2014. The next most common test location was a natural field around 2007/2008. The most common test location does change over time, and it has since decreased drastically as we reached 2020.

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.

```
#creatign bar graph of endpoint counts
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar()
```



```
which.max(table(Neonics$Endpoint))
```

```
## NOEL
##   25
```

```
##NOEL
##25
```

Answer: NOEL and 25 were the two most common endpoints. They are defined by

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) #checking class of collectDate
```

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

```
## [1] "factor" - is not recognized as a date

Litter$collectDate <- as.Date(Litter$collectDate) #changing it to date class
class(Litter$collectDate) #confirming new class of collectDate
```

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

```
## [1] "Date" - now recognized as a date

unique(Litter$collectDate) #checking whihc dates litter was sampled in Aug 2018
```

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

```
## [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`?

```
unique(Litter$plotID) #checking how many plots sampled at Niwot Ridge with 'unique'
```

```
##  [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
```

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

summary(Litter$plotID) #comparing it against 'summary' results
```

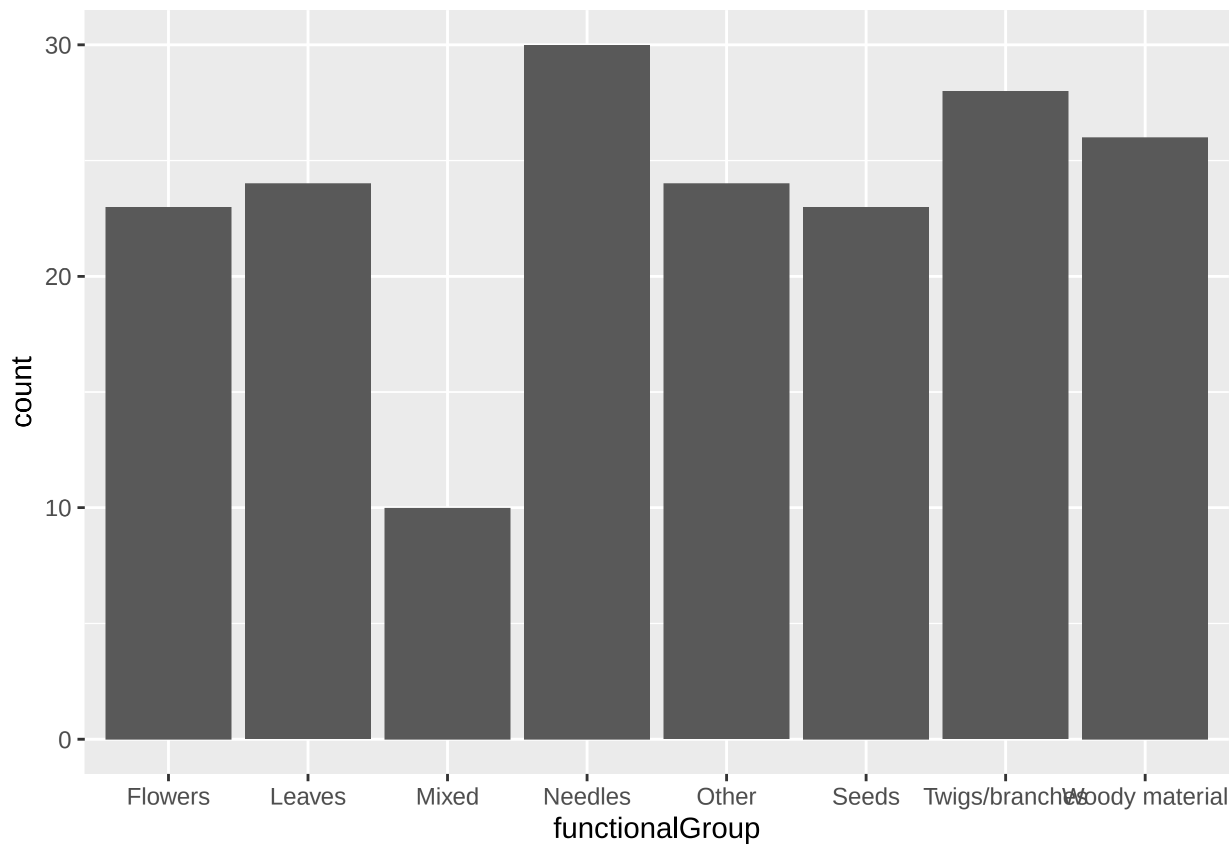
```
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061
##       20       19       18       15       14        8       16       17
## NIWO_062 NIWO_063 NIWO_064 NIWO_067
##       14       14       16       17
```

```
## [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051 NIWO_058
## [10] NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 ... NIWO_067
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061 NIWO_062 NIWO_063
##      20       19       18       15       14        8       16       17       14       14
## NIWO_064 NIWO_067
##      16       17
```

Answer: Both functions showed that 12 plots were sampled at Niwot Ridge. The summary function shows us how many tests were done at each of the 12 plots, giving us an overview of the whole dataset. The unique function only lays out what the 12 plots sampled are, without the count of how many tests were done at each plot.

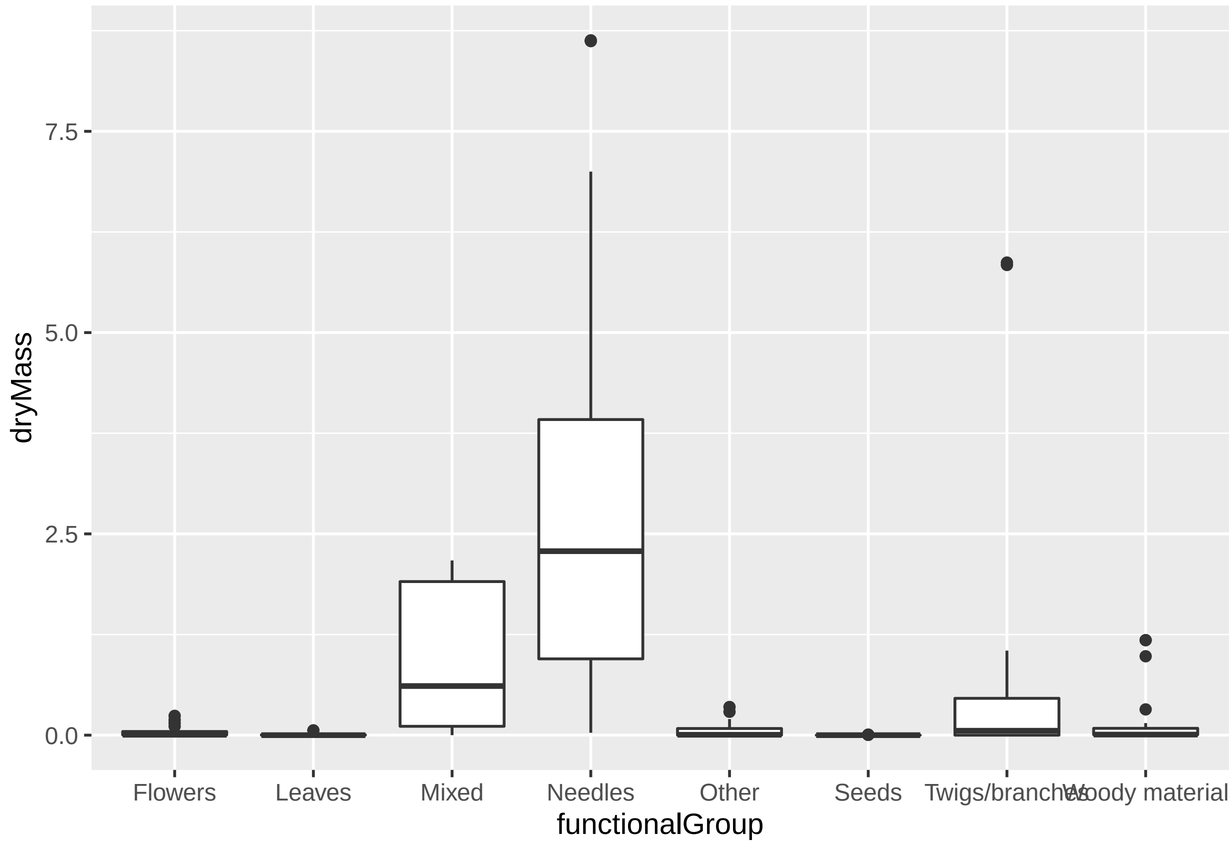
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#creating bar graph to show functional group counts
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.

```
#creating boxplot of the relation between dry mass and functional group
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass))
```

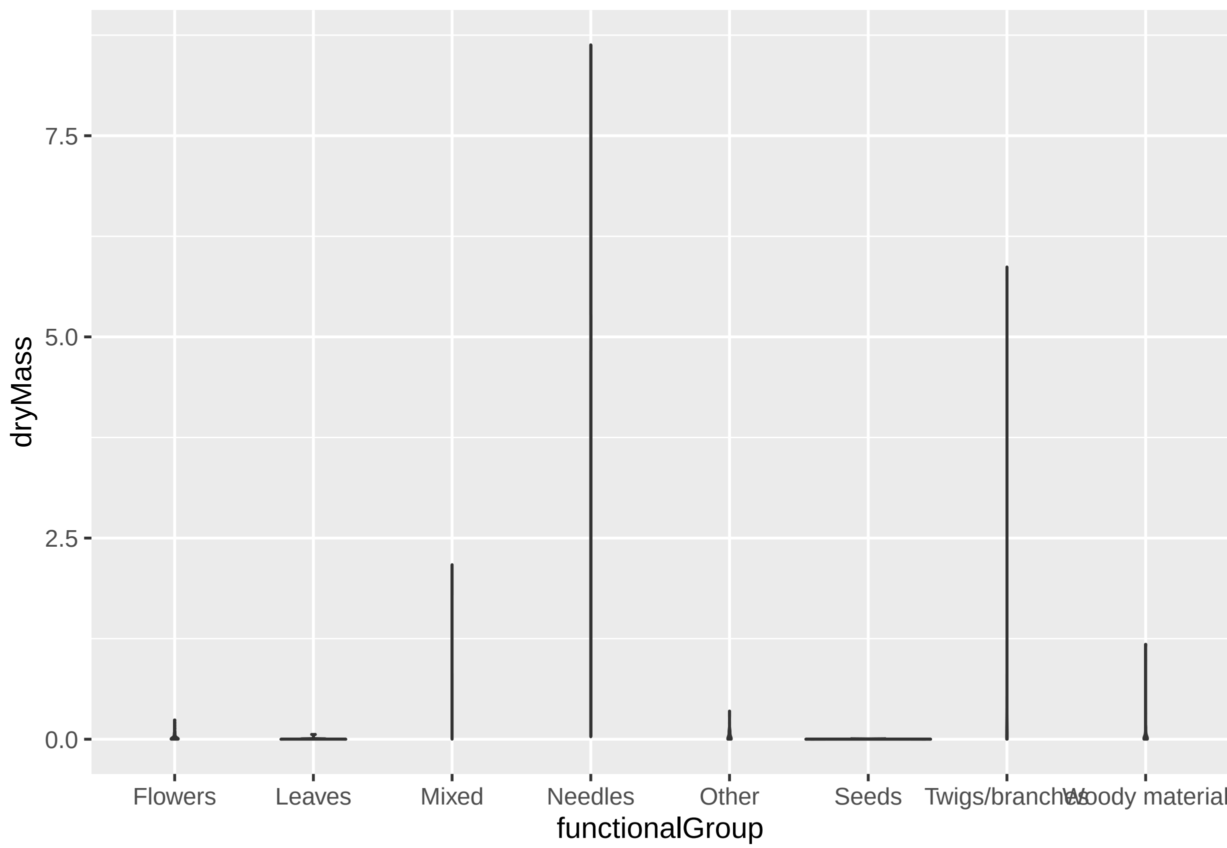



```r
#creating violin plot of the relation between dry mass and functional group
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
              draw_quantiles = c(0.25, 0.5, 0.75))
```

```
## Warning in regularize.values(x, y, ties, missing(ties), 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: The boxplot is easier to read, and provides a clearer picture of the data. It shows details like outliers and the median of the data. The violin plot only shows the complete range of the data, with no additional details about where all the points fall within that range.

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, with a greater range and maximum than other types of litter.