Assignment 3: Data Exploration

Jack Carpenter, Section #2 (Tuesday)

OVERVIEW

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

Directions

1. Change "Student Name, Section #" on line 3 (above) with your name and section number.
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., "FirstLast_A03_DataExploration.Rmd") prior to submission.

The completed exercise is due on <>.

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 add the `stringsAsFactors = TRUE` parameter to the function when reading in the CSV files.**

```
getwd()
```

```
## [1] "/Users/Jack/Documents/Duke/Spring 2022/Environmental Data Analytics/Environmental_Data_Analytics
```

```
setwd("/Users/Jack/Documents/Duke/Spring 2022/Environmental Data Analytics/Environmental_Data_Analytics_
getwd()
```

```
## [1] "/Users/Jack/Documents/Duke/Spring 2022/Environmental Data Analytics/Environmental_Data_Analytics
```

```
library(tidyverse)

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 ecotoxicologoy of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: Insects are agricultural pests, and controlling them with poisons on our food supply could have disastrous effects, so there is large incentive to study just how these insecticides impact

insect populations both for financial reasons and for human health reasons.

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 debris are a key factor in the recycling of nutrients in a forest system and their breakdown and reconsumption are key aspects of maintaing healthy soils. Woody litter and debris also are key fuels in fire-prone areas, like mountainous Colorado, and information on fuel loading rates of litter and debris is important for fire management.

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: * Sampling traps are locate both above and on the ground in pairs * Traps are within 20m x 20m or 40m x 40m sampling plots * Ground traps are sampled annually, elevated traps are sampled biweekly with deciduous vegetation and annually with evergreen vegetation (based on the location, these elevated traps are likely sampled annually)

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

```
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 studied are population and mortality. These effects are indicative of how strong/deadly an insecticide is and how well it kills insects compared to other behavioural changes.

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
```

Answer: The 6 most common species are honey bee, parasitic wasp, buff tailed bumblebee, carniolan honey bee, bumble bee, and italian honeybee. 5 of these 6 species are pollinators and decimating pollinator populations is bad for everyone. The other, parasitic wasp, is potentially a predatory control species that may also benefit agriculture by limiting pest populations. In short, all 6 top species are "farmers' friend" 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?

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

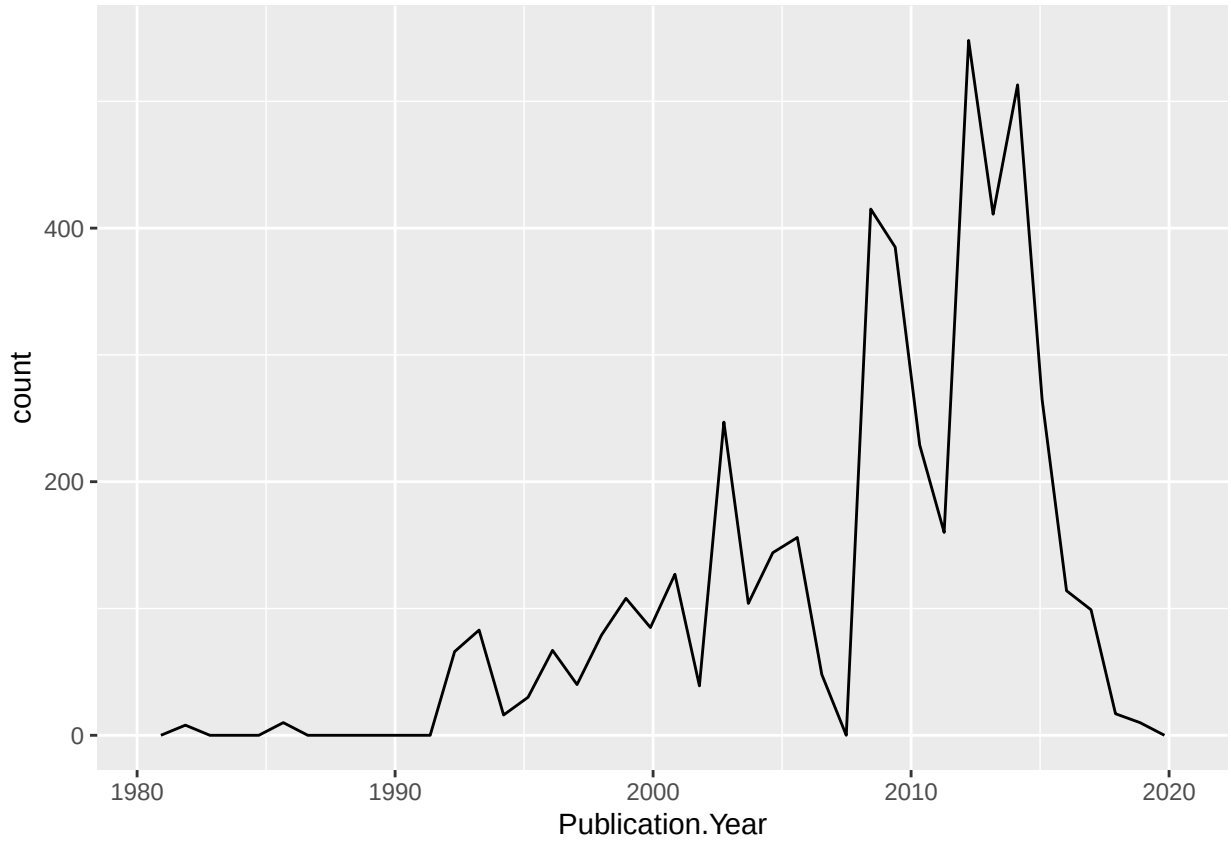
```
## [1] "factor"
```

Answer: The class of 'Conc.1..Author" is factor instead of numeric because there are several entries that are "NR" instead of numerical entries. As a result, R cannot read the whole column as numerical because some of the data is not entered as a number.

Explore your data graphically (Neonics)

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

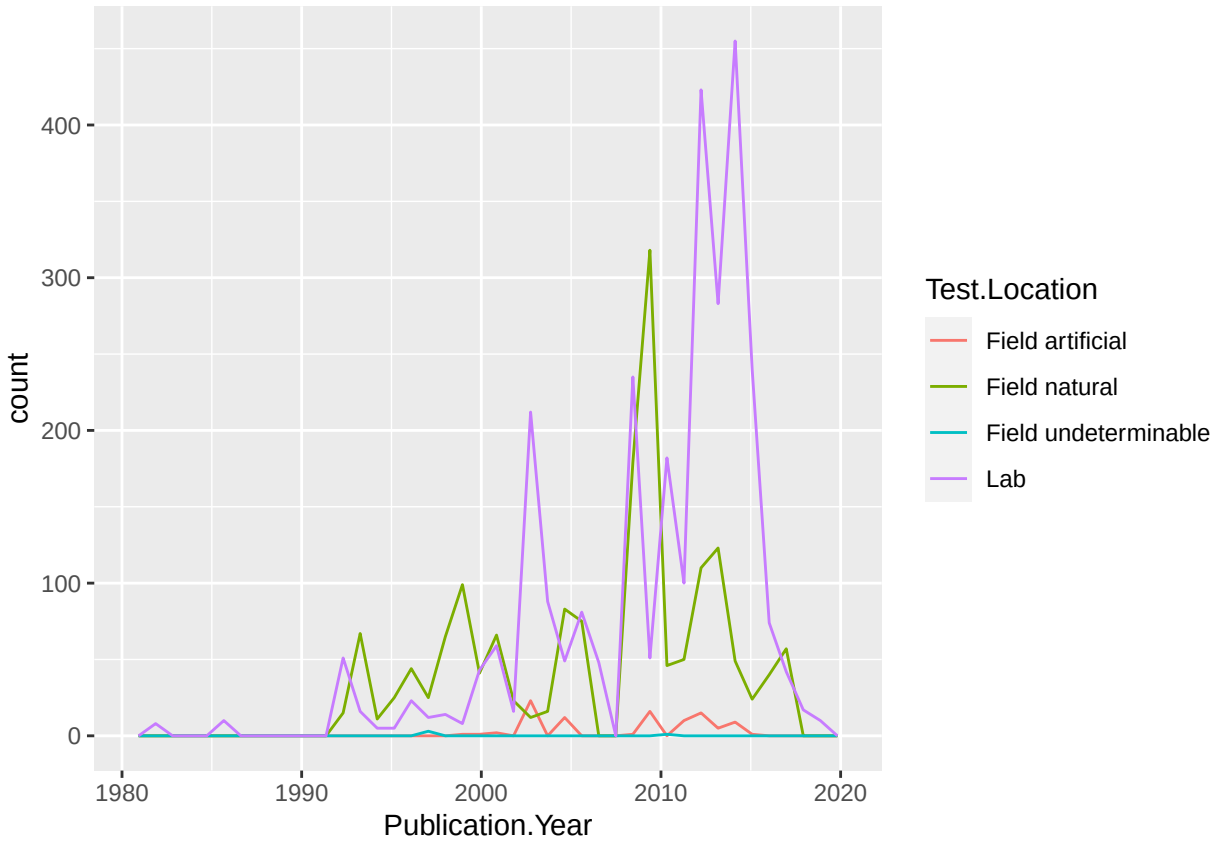
```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 40)
```




```
#how do I decide how many bins to use?
#I went with 40 because there is nearly 40 years' worth of data
```

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)
```

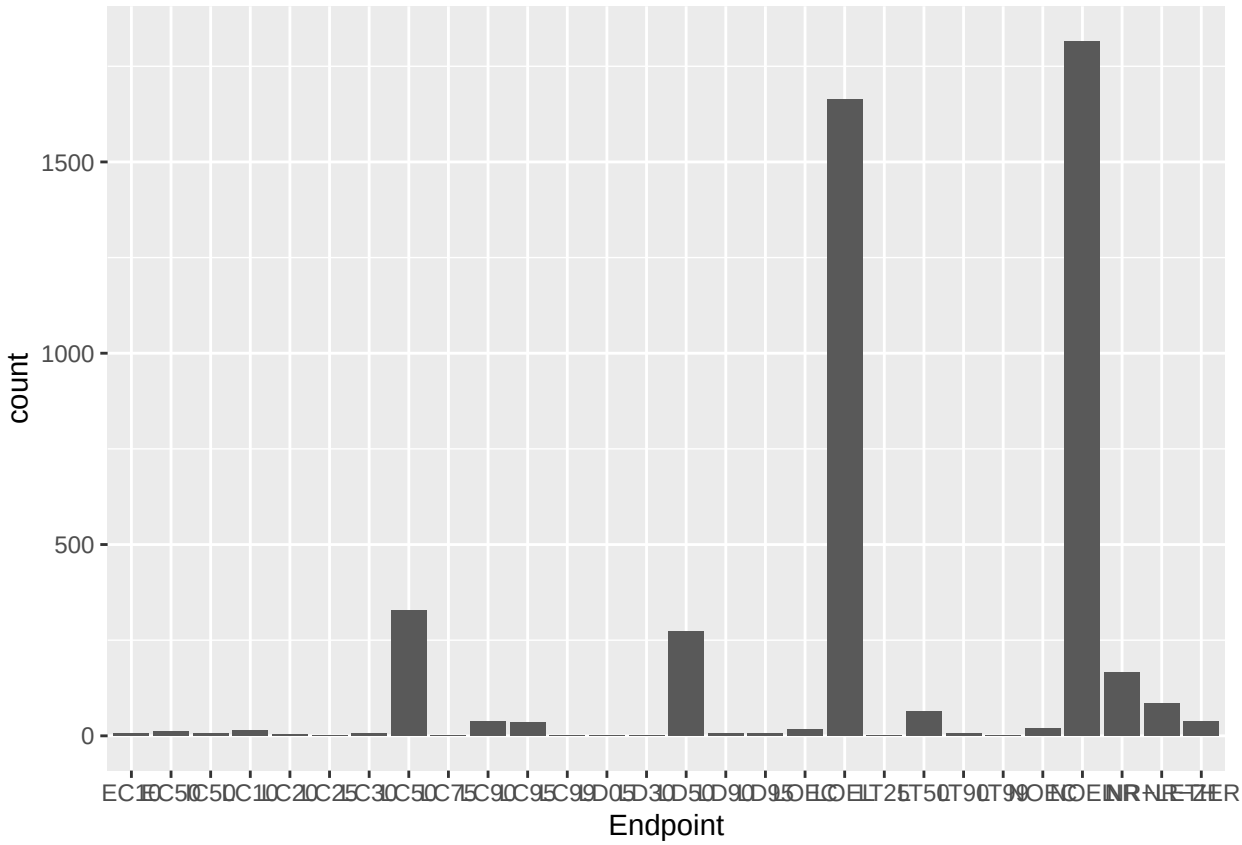


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common location appears to be the lab, but there are some years that a natural field is most common. Use of both the lab and natural field locations increase tremendously over time.

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.

```
#summary(Neonics$Endpoint)
#what do we have? How many bins to use?
#commmented out because we don't need bins with this one
ggplot(Neonics) +
  geom_bar(aes(x = Endpoint))
```



Answer: The two most common end points are NOEL and LOEL, respectively 'no observable effect level' and 'low observable effect level'. NOEL means that no statistically significant impacts were observed at the highest dose level, where LOEL means that statistically significant impacts were observed at the lowest dose level. Effectively, the two most common endpoints are opposite reactions.

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"
```

```
#comes back as factor, so need to change it
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
#should be a date now
class(Litter$collectDate)
```

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

```
#doublecheck
unique(Litter$collectDate)
```

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

```
#sampled 2018-08-02 and 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`?

```
help("unique")
#seeing what exactly it does
unique(Litter$siteID)
```

```
## [1] NIWO
## Levels: NIWO
```

```
#only NIWO comes up - all plots were sampled at Niwot Ridge
summary(Litter$siteID)
```

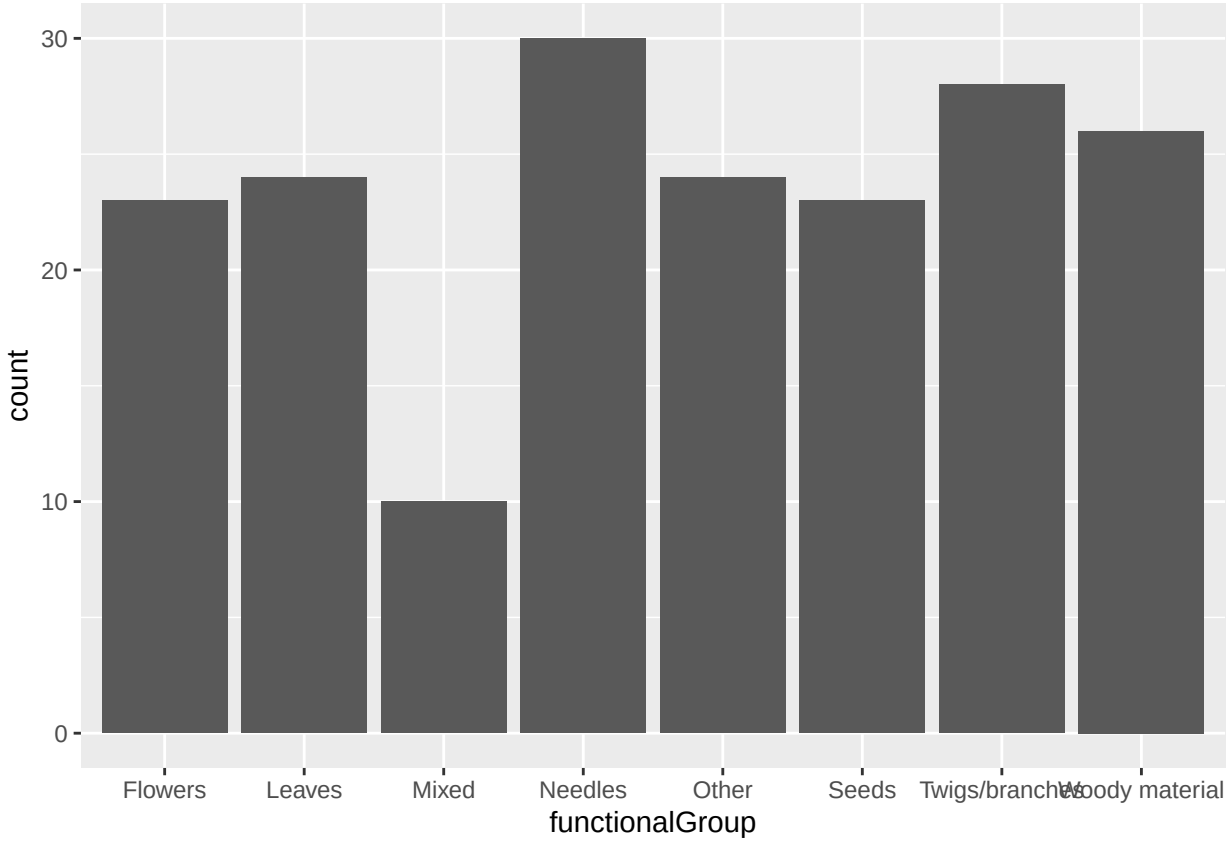
```
## NIWO
##  188
```

```
#188 sites
```

Answer: Unique looks for distinct data, where summary shows how many data points are in the column. In this case, unique only pulls up one site because the plots are all sampled at the same site, and summary pulls up the number of sites listed in the column. Combined, we gain that all 188 sites are niwot ridge.

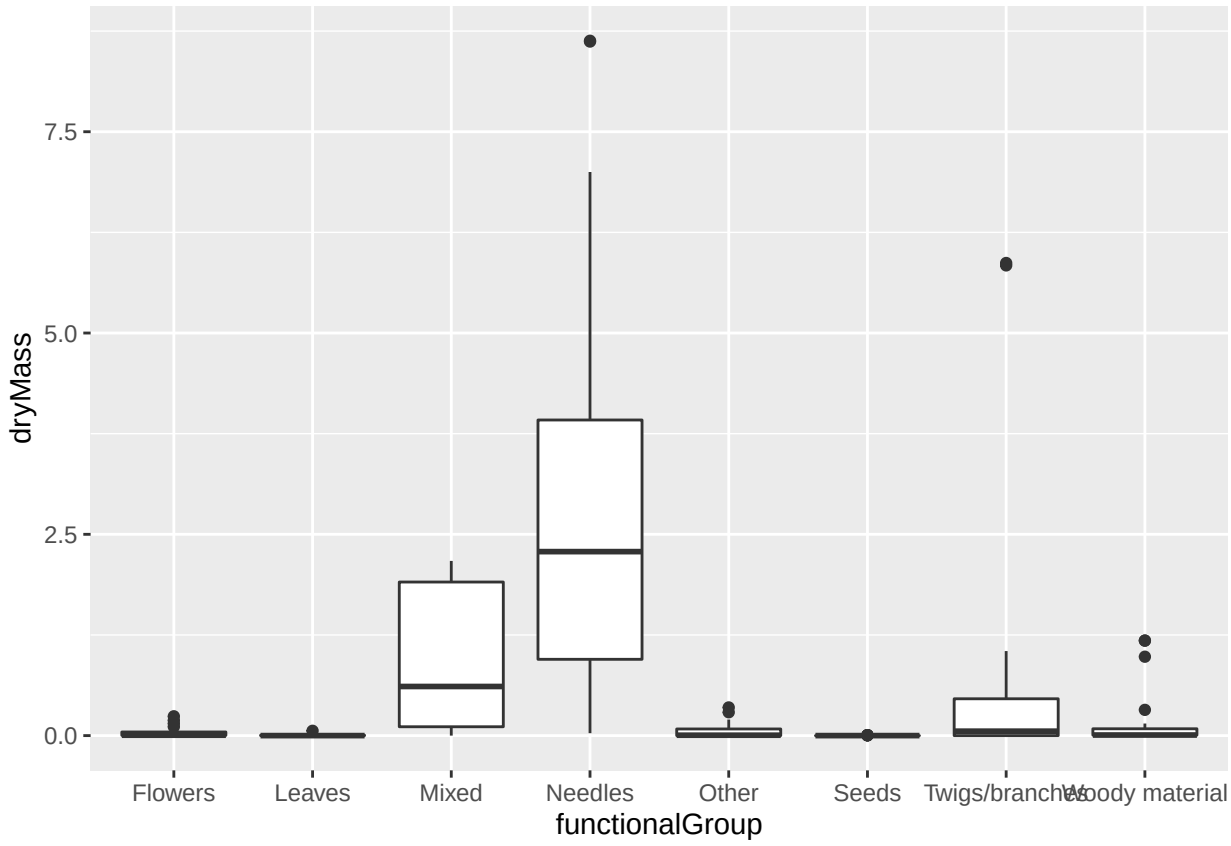
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

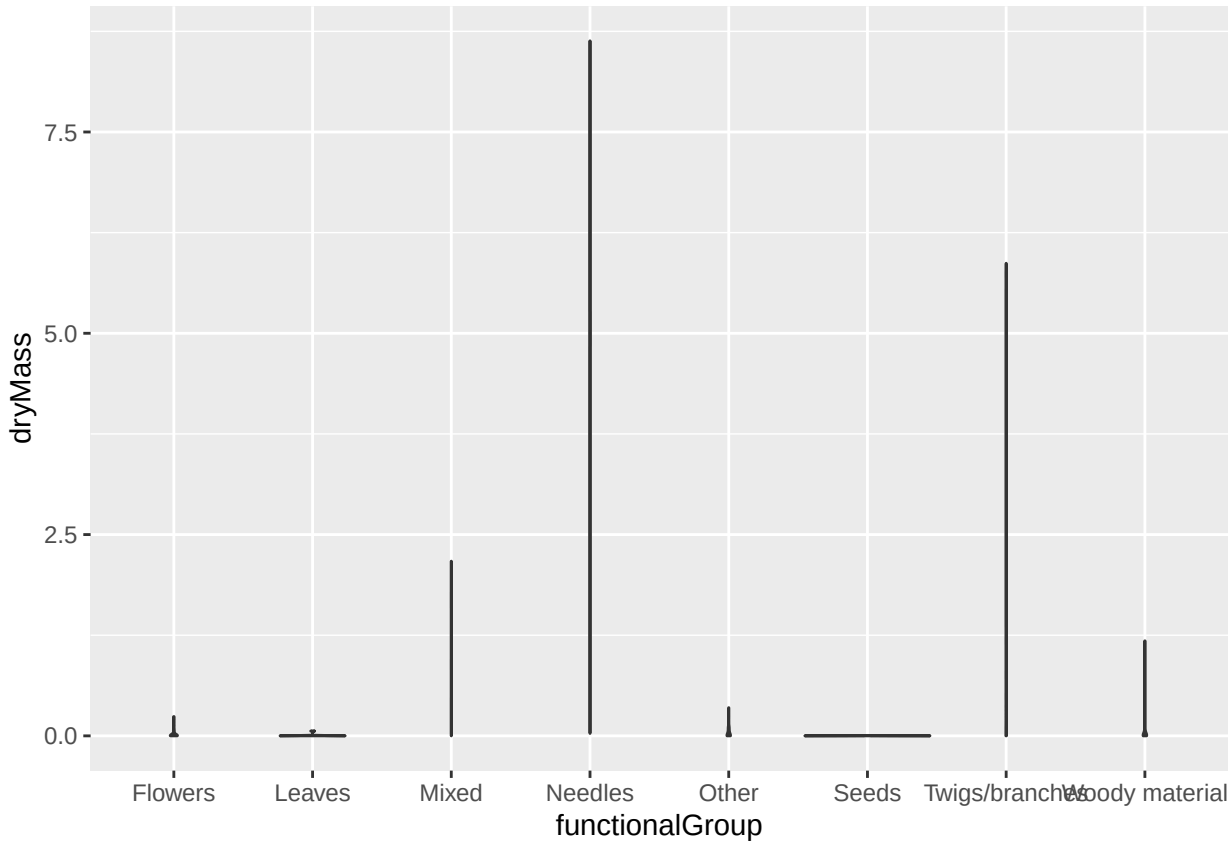


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




```
#now a violin plot
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass))
```



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

> Answer: The boxplot shows the statistical spread better, where the violin plot only shows a single line for each functional group.There doesn't seem to be enough data points for the violin plot to be properly fleshed out.

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

> Answer: Needles by far have the highest biomass (makes sense, it is a pine forest). After that, mixed has the highest biomass with a mean at about 0.75. The other categories are all hard to distinguish on this scale, but all are close to zero, except for twigs/branches with an upper quartile around 0.75.