Assignment 3: Data Exploration

Curtis Cha, Section #2

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.**

```
library(tidyverse)
setwd("C:/Users/curtx/Desktop/Environnmental Data Analytic/Environmental_Data_Analytics_2022/Data/Raw/")
Neonics<- read.csv("ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = T)
Litter <- read.csv("NEON_NIWO_Litter_massdata_2018-08_raw.csv", stringsAsFactors = T)
```

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: Insecticides are used in agriculture to protect crops for consumption or processing. It's important for farmer's to maintain their livelihoods and provide communities with the foods and materials needed for survival or development. However, the impact of insecticides can't be completely controlled, and non-pest species (both insect and animal). Bees, which provide important ecosystem services are particulary vulnerable to neonicotinoids. It's important to understand why neonicotinoids are used, how they are used, and the impacts on different insect groups.

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 Wood Debris serve important roles as nutrient sources for decomposers, which break down organic material further. This organic material continues to serve as nutrient sources for plant life. It would be important to observe how much leaf/wood litter there are in a given space, and what types of litter there are.

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: * First, sample plots are taken from areas with $>$ 2 m. vegetation. Of the tower plots within these areas, sample plots are taken based on random selection of these tower plots. * These sample plots are 40x40 m in size, but tower plots with low-statured vegetation have additional 20x20m plots. These smaller 20x20m plots also have litter traps included within its area. * For each area of 400 m^2, there are two litter traps (elevated and ground). The ground traps are sampled once a year while the eleveated traps' sample frequency depends depending on the surrounding forest type (More frequent sampling for deciduous forest types).

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
cat("Number of rows/cases: ", as.character(dim(Neonics)[1]) )
```

```
## Number of rows/cases:  4623
```

```
cat("\nNumber of columns/attributes ", as.character(dim(Neonics)[2]) )
```

```
##
## Number of columns/attributes  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 of the toxin on insects are Mortality and Population. Studying this column would be important since it discusses the effect of the toxin on insect species. It could provide insight on the effectiveness of the toxin in exterminating pests and the harmful impacts on non-pest insect species. The effect column itself describes what the measurements tell us. In this case, the most common effects measured on mortality and population.

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 six most common species studied (excluding the Other category) are the Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and the Italian Honey Bee. This is informative because we can observe the impacts of the toxin on non-pest species. Since bees provide important ecosystem services through pollination of flowers and plants, it's important to determine if the toxin is overall useful for crop protection. The toxin could protect

crops from pests but also risk crop failure due to lack of bee pollination.

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 the given attribute is factor. It is not numeric because these numbers cannot be compared to each other. The following column Conc.1..Units.Author show the units ofthe concentrations. Since different cases use different units, the concentration numeric values should not be numeric type.

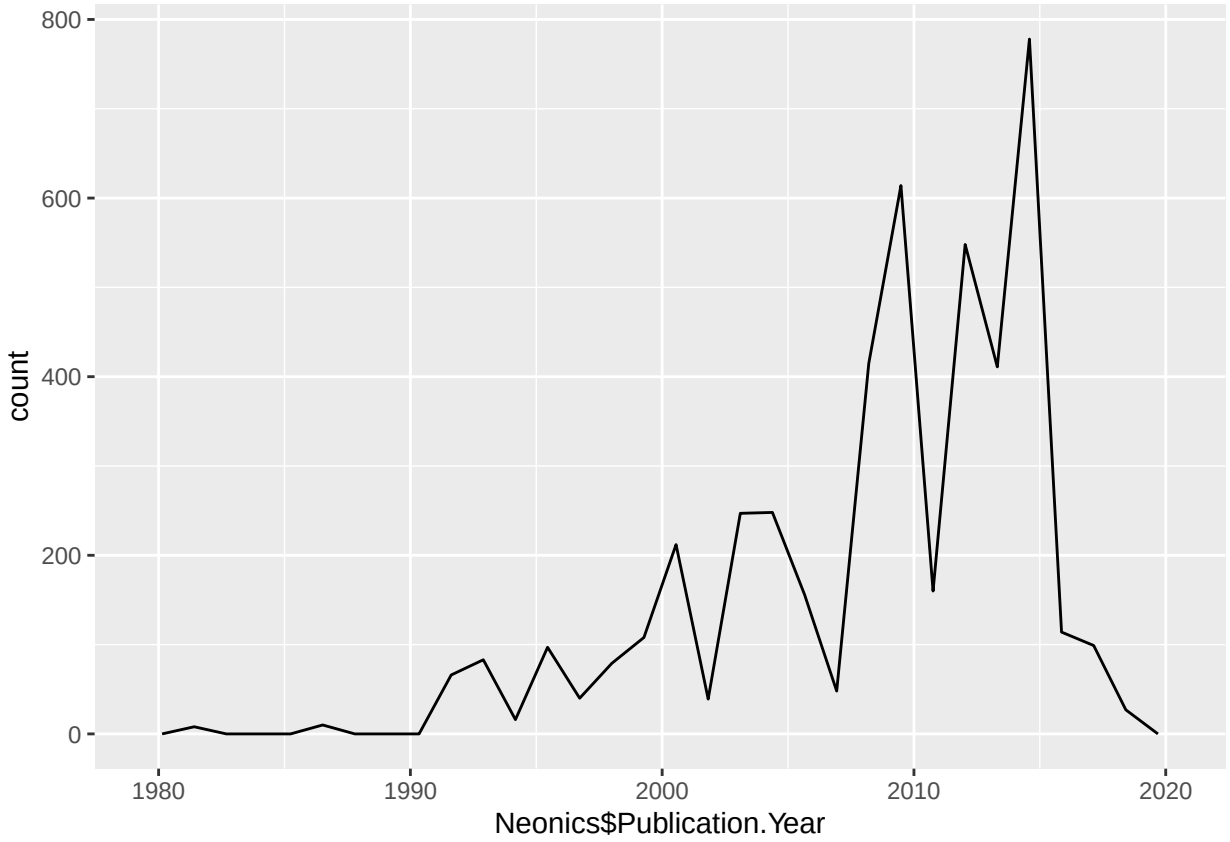
Explore your data graphically (Neonics)

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

```
ggplot(data = Neonics, aes(x = Neonics$Publication.Year)) + geom_freqpoly()
```

```
## Warning: Use of `Neonics$Publication.Year` is discouraged. Use
## `Publication.Year` instead.
```

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



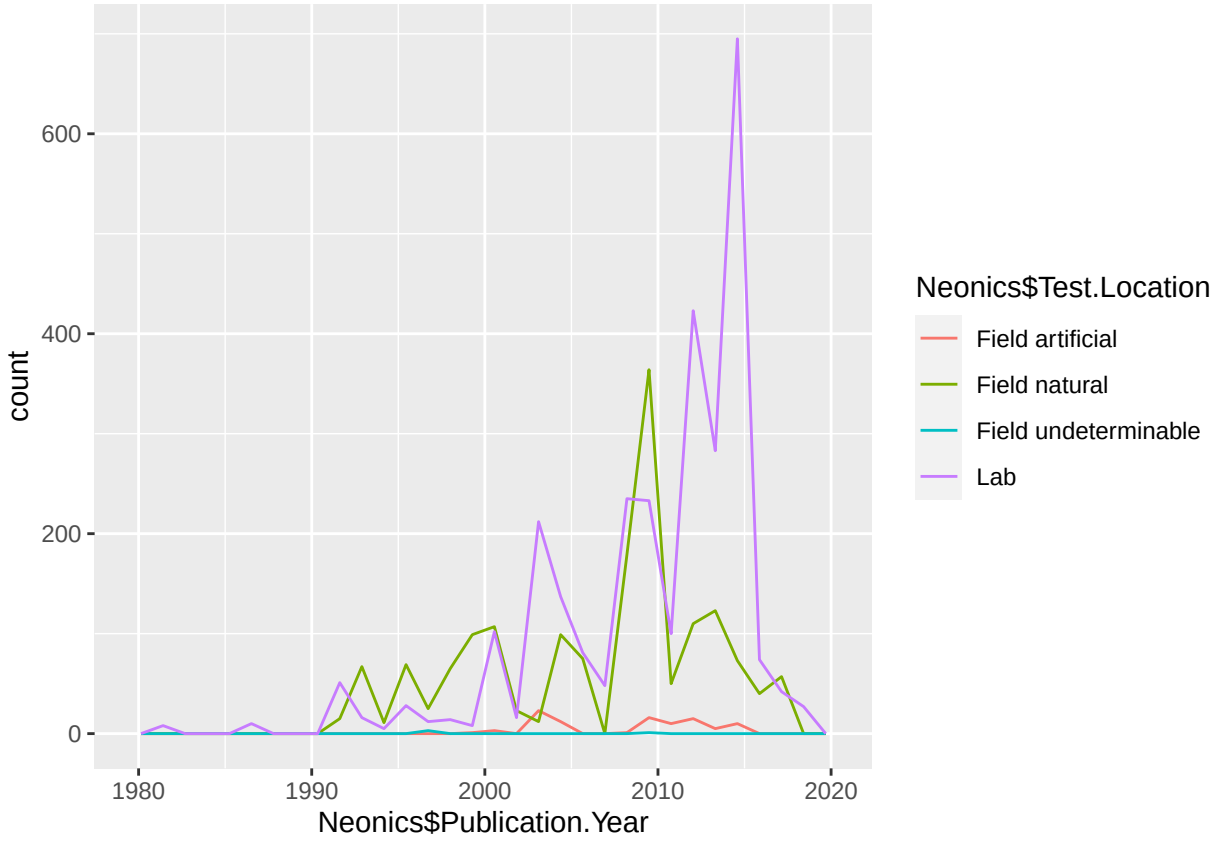
10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
ggplot(data = Neonics, aes(x = Neonics$Publication.Year, colour = Neonics$Test.Location)) + geom_freqpol
```

```
## Warning: Use of `Neonics$Publication.Year` is discouraged. Use
## `Publication.Year` instead.
```

```
## Warning: Use of `Neonics$Test.Location` is discouraged. Use `Test.Location`
## instead.
```

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



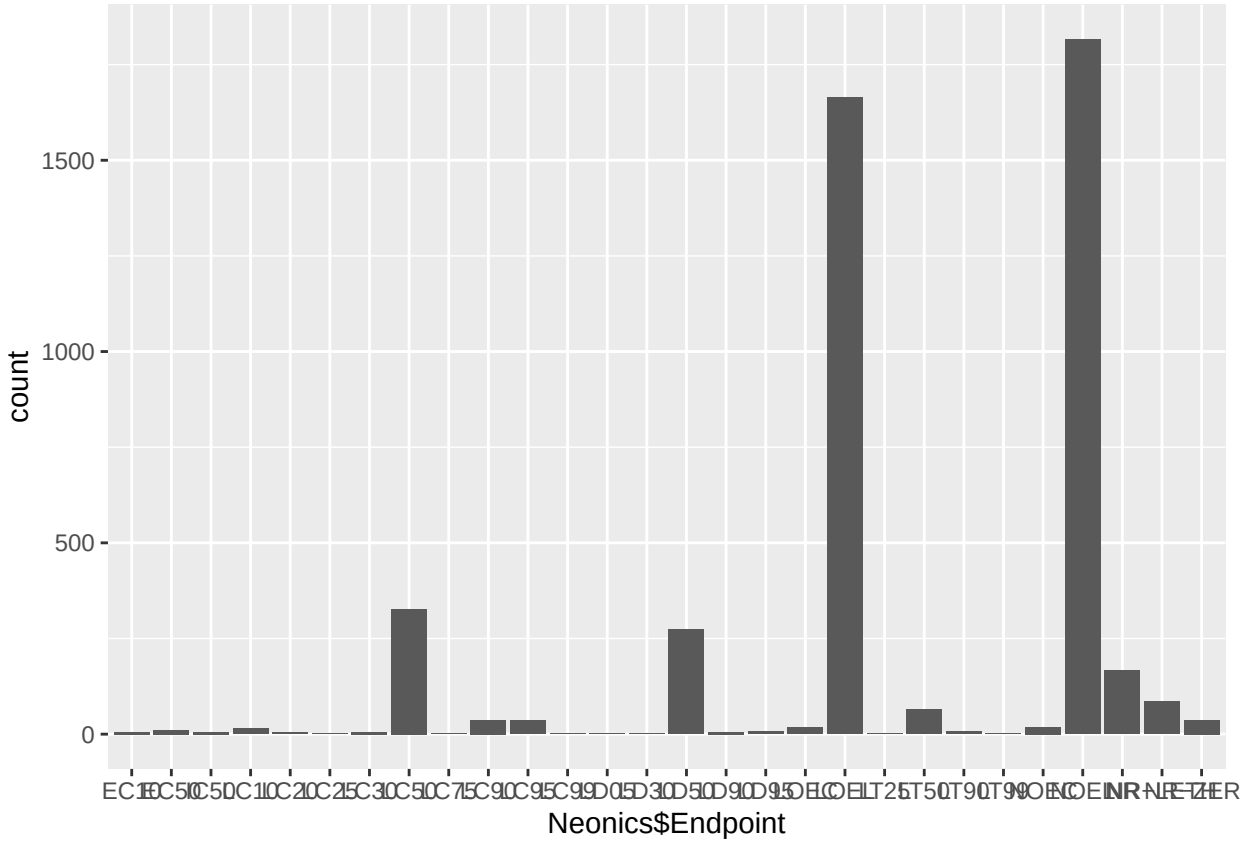
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most commmon test location are lab and field (natural). Around 2010, the field natural test locations were more common than lab, but as time moved one, lab become more common again.

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(data = Neonics, aes(x = Neonics$Endpoint)) + geom_bar()
```

```
## Warning: Use of `Neonics$Endpoint` is discouraged. Use `Endpoint` instead.
```




```
summary(Neonics$Endpoint)
```

```
##    EC10    EC50    IC50    LC10    LC20    LC25    LC30    LC50    LC75    LC90
##       6      11       6      15       5       1       6     327       1      37
##    LC95    LC99    LD05    LD30    LD50    LD90    LD95    LOEC    LOEL    LT25
##      36       2       1       1     274       6       7      17    1664       1
##    LT50    LT90    LT99    NOEC    NOEL      NR NR-LETH NR-ZERO
##      65       7       2      19    1816     167      86      37
```

Answer: The most common endpoint tpyes are LOEL and NOEL. LOEL stands for lowest observable effect level (lowest dose) while NOEl stands for no observable effect level (highest dose). Both endpoint codes represent either high or low doses that produe significantly different effects than that of 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"
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
```

```
##
##     date, intersect, setdiff, union
```

```
Litter$collectDate <- ymd(Litter$collectDate)
unique(Litter$collectDate)
```

```
## [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)
```

```
##  [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
```

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

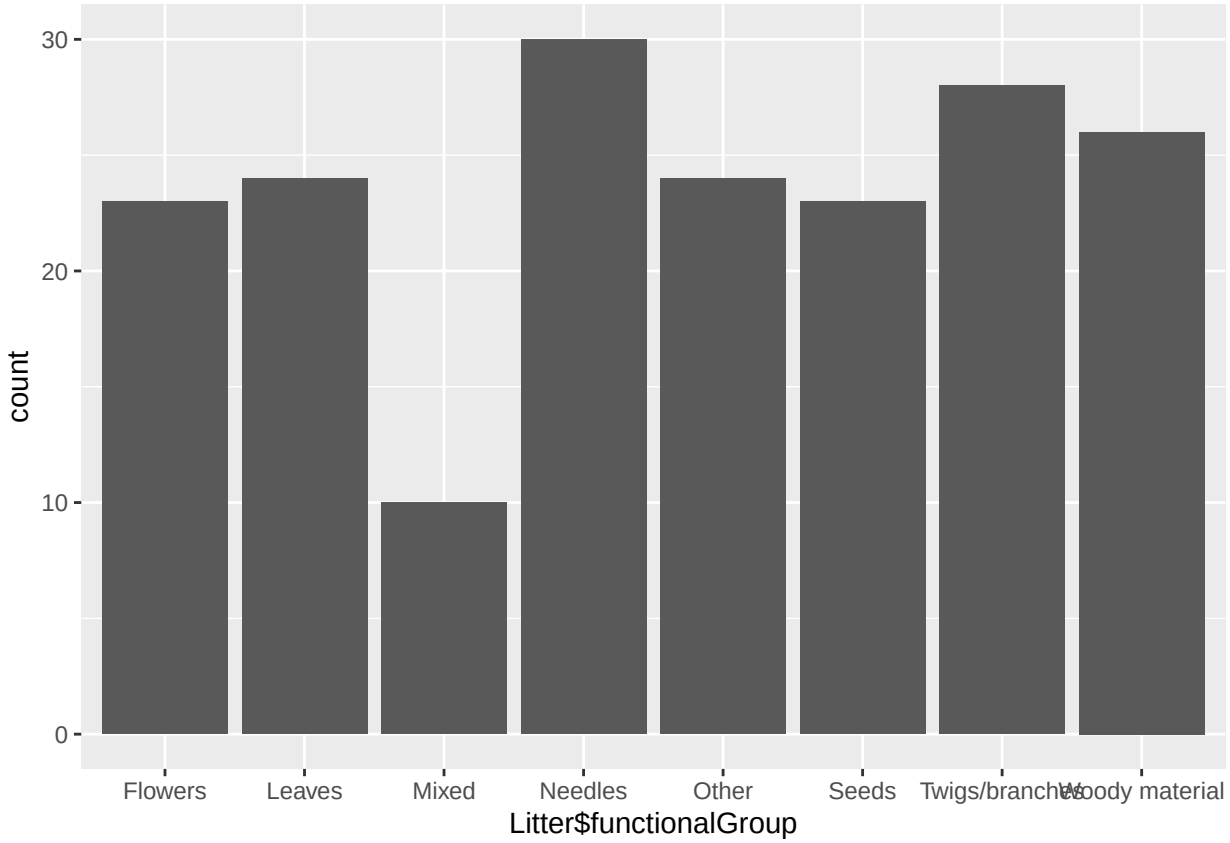
```
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061
##       20       19       18       15       14        8       16       17
## NIWO_062 NIWO_063 NIWO_064 NIWO_067
##       14       14       16       17
```

Answer: There are 12 unique sites sampled at Niwot Ridge. In contrast to summary(), the unique function only gives out the unique values in a given column. The summary of the same column would provide the number of rows that share the same column value. With summary(), I could determine from which plot were the most and least samples taken. With unique(), I cannot determine that

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.

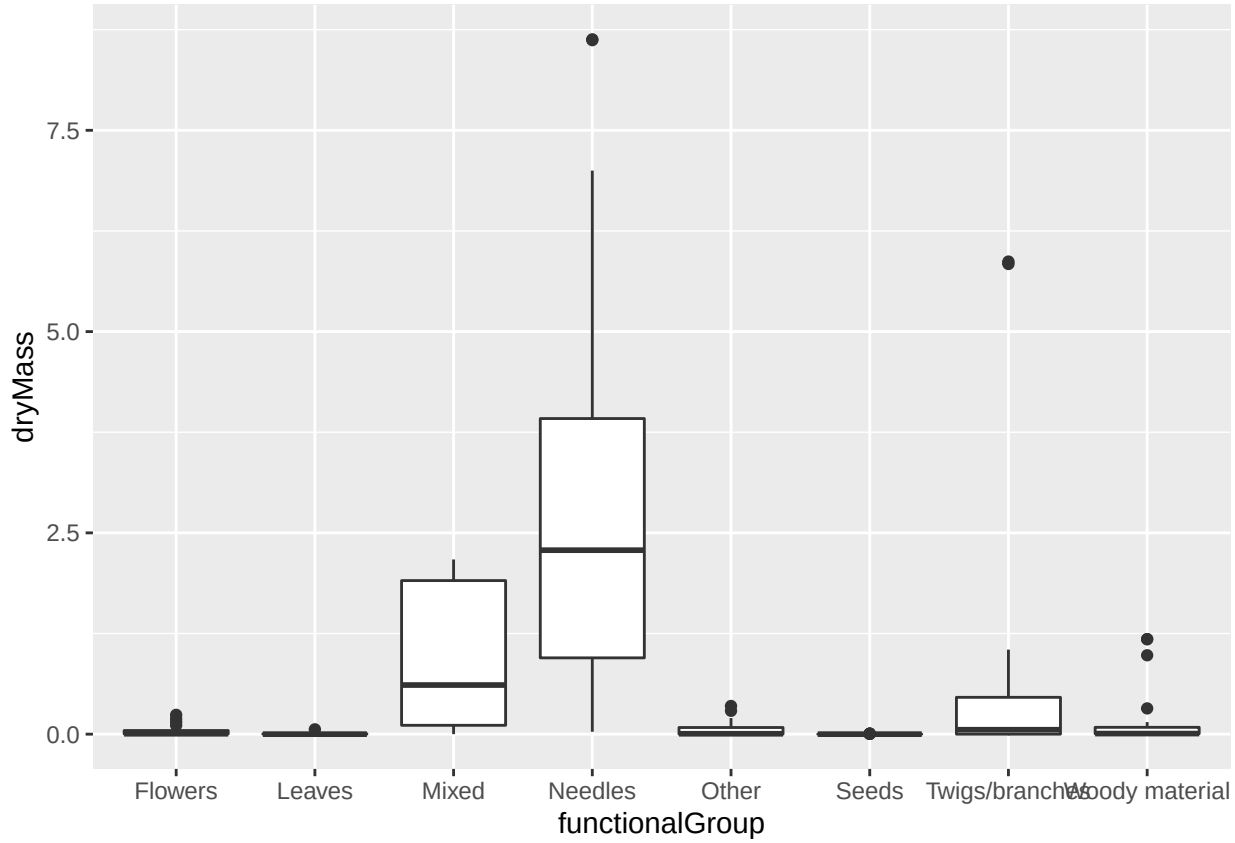
```
ggplot(data = Litter, aes(x = Litter$functionalGroup)) + geom_bar()
```

```
## Warning: Use of `Litter$functionalGroup` is discouraged. Use `functionalGroup`
## instead.
```

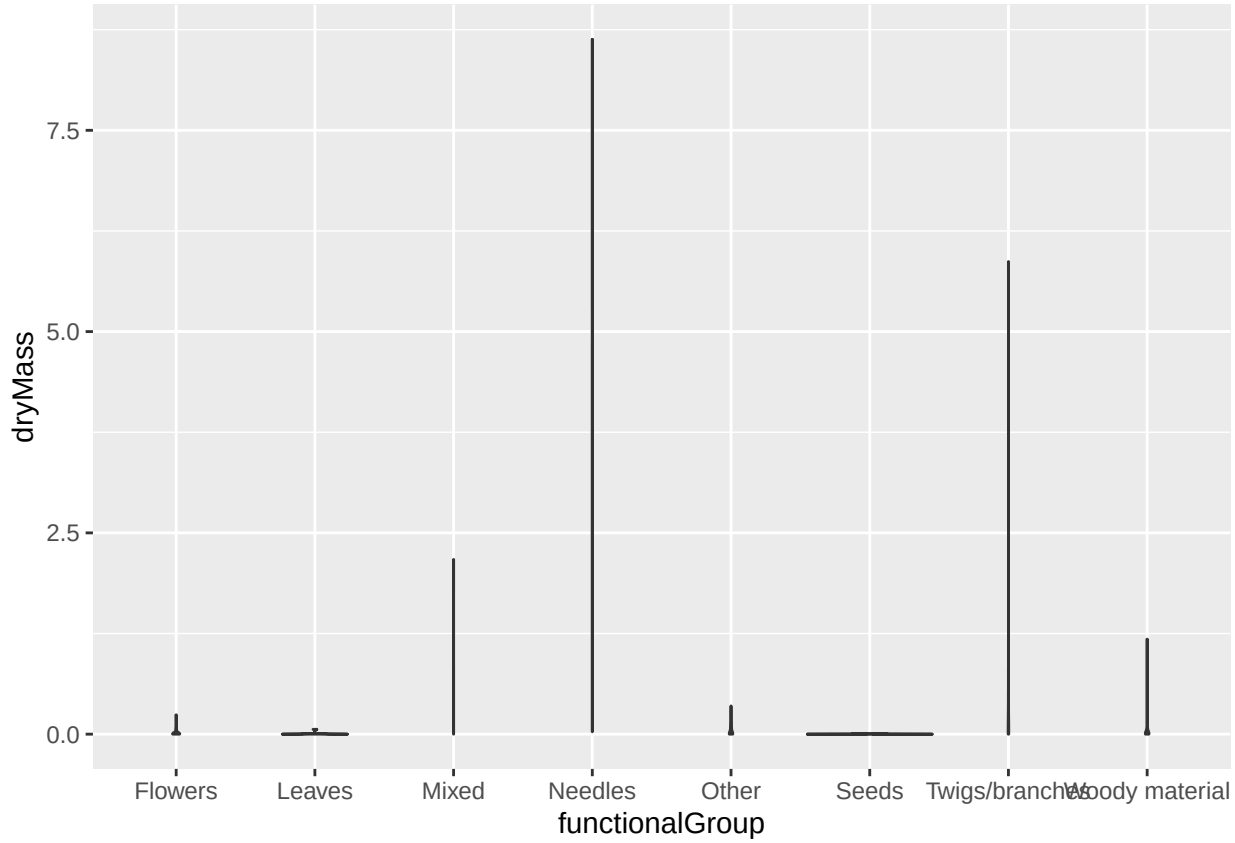


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




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



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

> Answer: The boxplot has better visualization because we can more clearly observe the quantile values. Since the geom_violin function also factors the density of a given value, the "density" of a quantile value depends the size of the data. When observing dry mass by functional groups, there aren't enough data in each functional group for a violin plot to be useful. Without enough data, a violin plot cannot display the quantiles or outliers well. The boxplot does not take into account the "data density" or "enough data", rather it just displays the quantile values and outliers. What type(s) of litter tend to have the highest biomass at these sites?
>
> Answer: The Mixed and Needles groups have the highest dry mass at the 12 sites. The mean and maximum values of dry mass of both are higher than the other groups' values.