Assignment 3: Data Exploration

GuruBandaa Khalsa

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.

```
#1. I checked my working directory, loaded tidyverse, and uploaded and named two datasets.
getwd()
```

```
## [1] "/Users/survivormangb/Desktop/Masters at Duke/Second Year/Fall Semester/Environmental Data Analy
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.2

## Warning: package 'ggplot2' was built under R version 4.1.2

## Warning: package 'tibble' was built under R version 4.1.2

## Warning: package 'tidyr' was built under R version 4.1.2

## Warning: package 'readr' was built under R version 4.1.2
```

```
## Warning: package 'dplyr' was built under R version 4.1.2

## Warning: package 'stringr' was built under R version 4.1.2

## Warning: package 'forcats' was built under R version 4.1.2
```

```
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: Neonicotinoids are very neurtoxic and negatively affect insects, especially insect larvae. It makes sense that we are interested in looking at the effects of insect exposure to neonicotinoids.

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 woody debris on forest floors are prime habitat for insects and insect larvae. There is plenty of cover from predators, insulation from cold temperatures, and decomposing matter that provides insects with food.

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 sampled at terrestrial NEON sites with woody vegetation that is >2 meters tall 2. Sampled in tower plots in 20 40 meter by 40 meter plots in forested tower airsheds; Sampled in 4 40 meter by 40 meter plots and 26 20 meter by 20 meter plots in areas with low-saturated vegetation over the tower airsheds 3. There are 1-4 litter trap pairs per plot: a pair includes one elevated and one ground trap deployed for every 400 square meter plot area

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#5. I used the 'dim' function to find the dimensions of the 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?

```
#6. I viewed the Neonics dataset and then used the 'summary' function on the "Effect" column to determi
View(Neonics)
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 appear to be mortality and population. These effects might be of interest because it would be important to know if insect exposure to neonicotinoids is commonly lethal or has significant impacts on insect population. This may help determine the severity of exposure and how much of a risk it poses to insects. Also, if concentration of neonicotinoids is known, the prevalence of insect mortality would indicate neonicotinoid toxicity level.

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.

```
#7. I used the 'summary' function to determine the six most commonly studied species in the dataset.
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 commonly studied species in the dataset are Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. All of these species belong to the order Hymenoptera. They might also be of interest over other insects because they are all pollinators. There existence is crucial for the pollination and continued life of many plants on Earth.

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
#8. I used the 'class' function to determine the class of Conc.1..Author in the Neonics dataset.
class(Neonics$Conc.1..Author.)
```

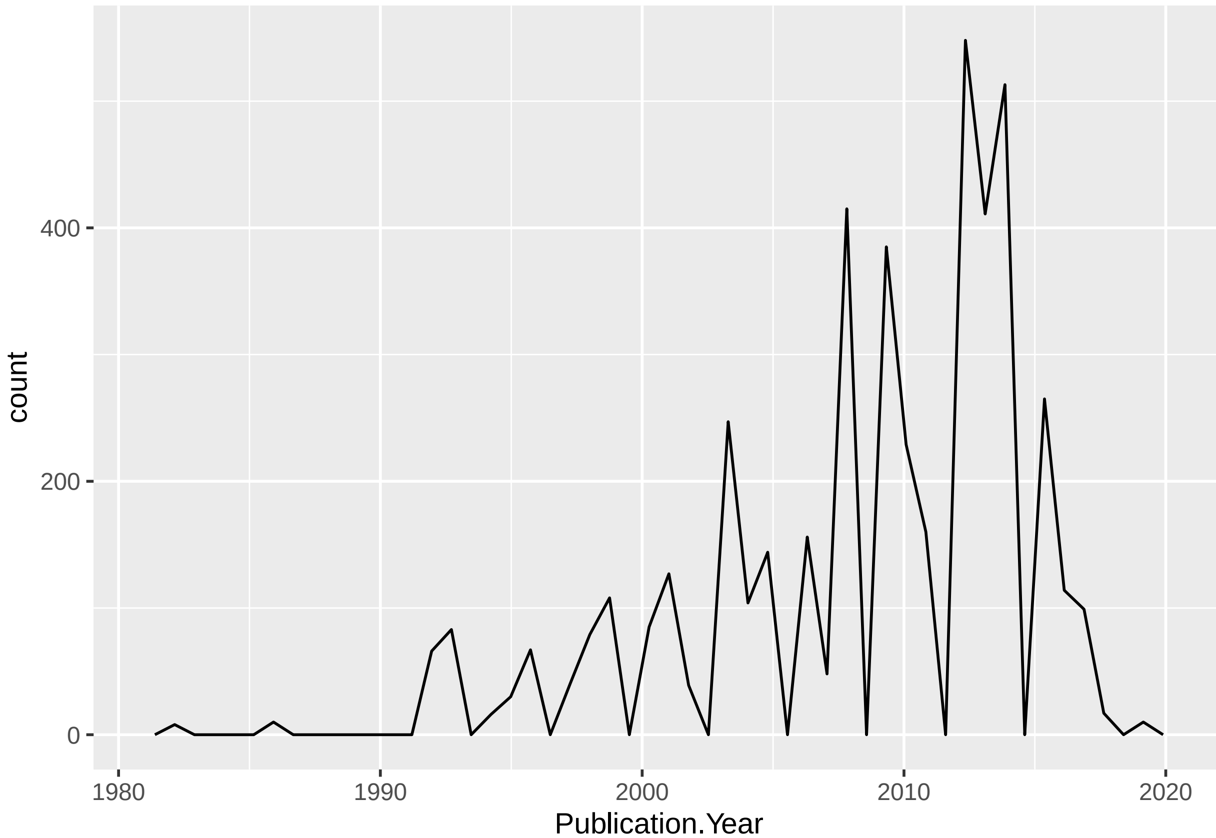
```
## [1] "factor"
```

Answer: Conc.1..Author is a factor. It is not numeric because even though the values are numeric, numeric variables can exist as a factor class. This enables them to be used in statistical modeling.

Explore your data graphically (Neonics)

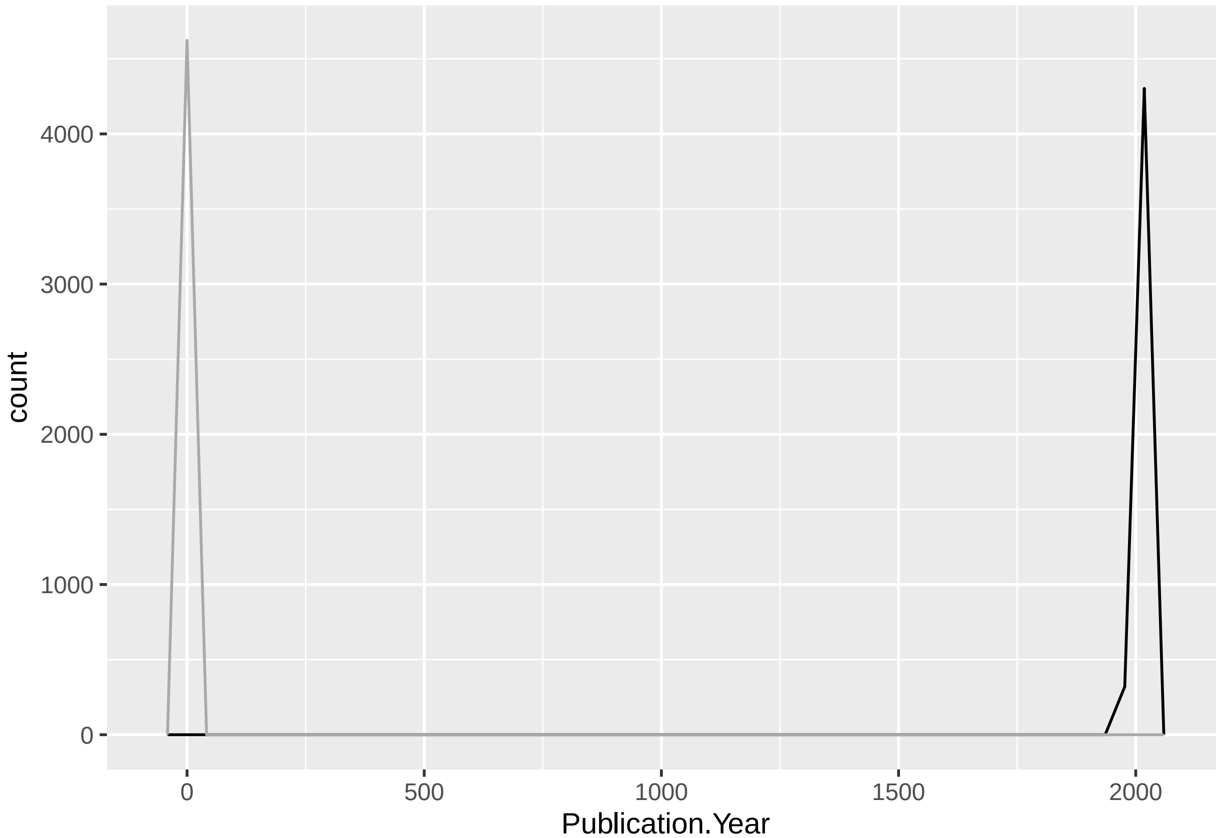
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```r
#9. I used 'geom_freqpoly' to generate a plot of the number of studies conducted by publication year.
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.

```
#10. I reproduced the same graph but used 'color' to add a color aesthetic for different Test.Location
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50) +
  geom_freqpoly(aes(x = as.numeric(Test.Location)), color = "darkgray", bins = 50)
```

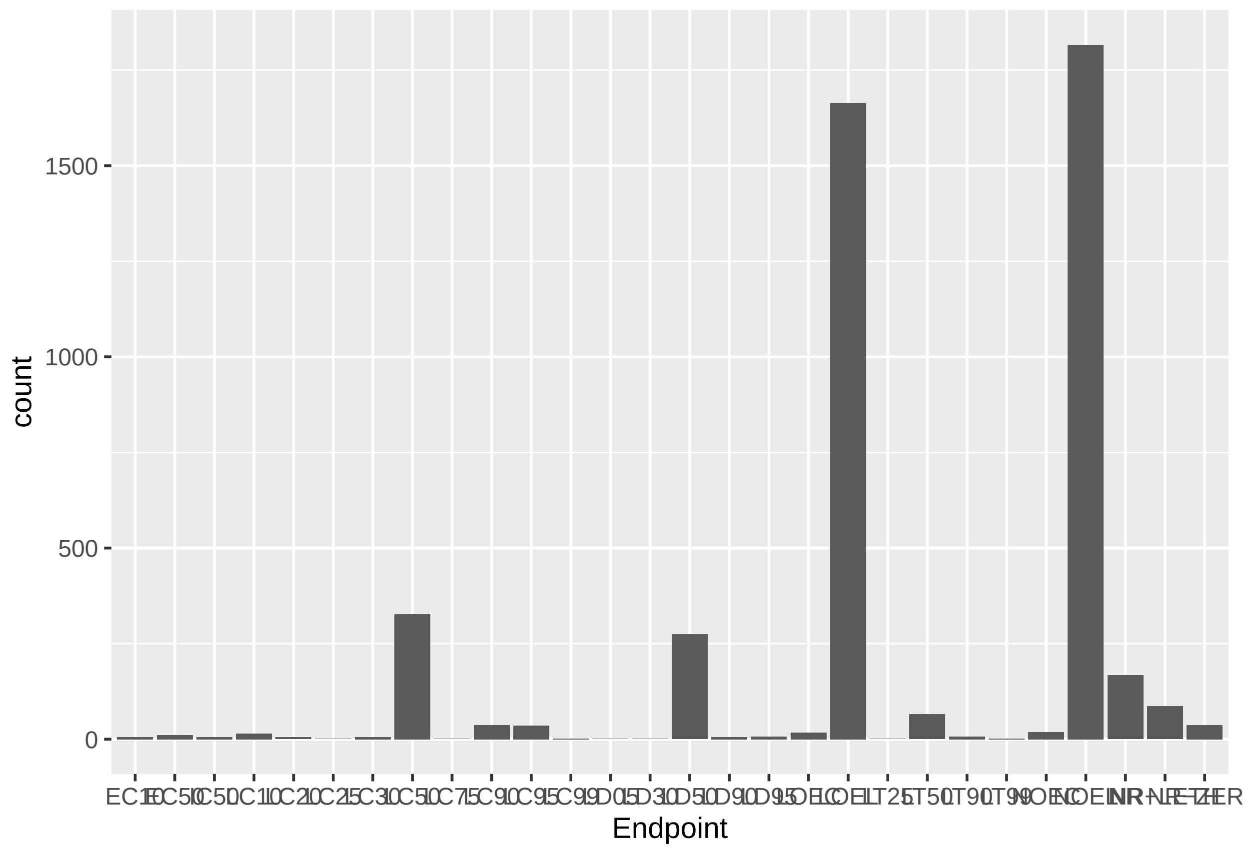


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are Lab and Field natural. They do differ over time. Lab is the most common test location in the early publication years while Field natural is the most common test location in the later publication years.

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.

```
#11. I used 'geom_bar' to create a bar graph of Endpoint counts.
ggplot(Neonics) +
  geom_bar(aes(x = Endpoint))
```



Answer: The two most common end points are LOEL and NOEL. They are they defined as Lowest Observable Effect Level and No Observable Effect Level.

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.

```
#12. I used the 'class' function to determine the class of collectDate.  It was not a date.  It was a f
class(Litter$collectDate)
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate, format = "%Y-%m-%d")
class(Litter$collectDate)
```

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

```
print(Litter$collectDate)
```

```
##   [1] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##   [6] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
```

```
##  [11] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [16] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [21] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [26] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [31] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [36] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [41] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [46] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [51] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [56] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [61] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [66] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [71] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [76] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [81] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [86] "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02" "2018-08-02"
##  [91] "2018-08-02" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
##  [96] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [101] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [106] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [111] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [116] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [121] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [126] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [131] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [136] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [141] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [146] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [151] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [156] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [161] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [166] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [171] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [176] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [181] "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30" "2018-08-30"
## [186] "2018-08-30" "2018-08-30" "2018-08-30"
```

```
unique(Litter$collectDate[2018-08])
```

```
## [1] NA
```

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`?

```
#13. I used the 'unique' function to determine how many plots were sampled at Niwot Ridge.
unique(Litter$plotID)
```

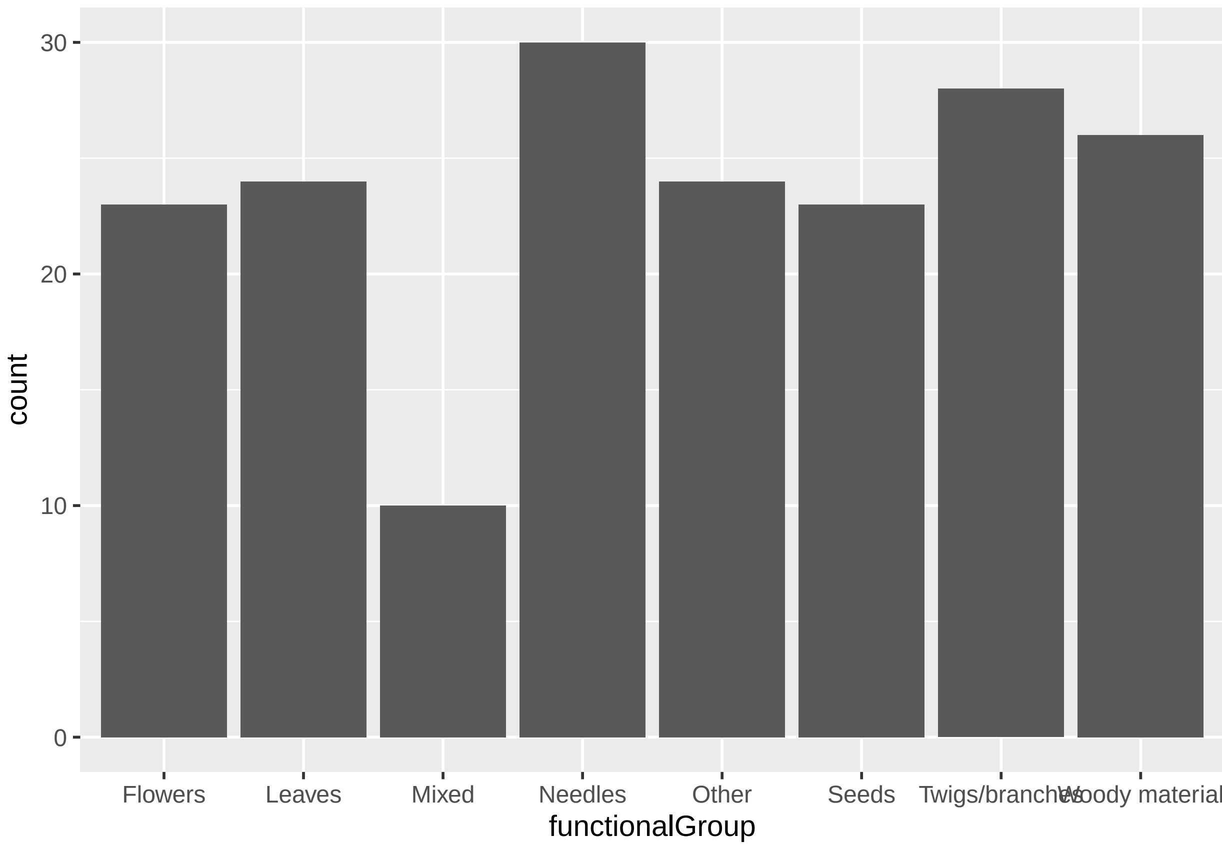
```
##  [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

```
plots <- unique(Litter$plotID)
#count(Litter$plotID)
```

Answer: The information is obtained from 'unique' different from that obtained from 'summary' because 'unique' is able to delete duplicate entries, whereas 'summary' does not do this.

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.

```
#14.
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.

```
#ggplot(Litter) +
#  geom_bar(aes(x = dryMass))

#geom_violin(aes(x = dryMass, functionalGroup
```

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

Answer:The boxplot is easier to read with this data. The data is displayed more logically.

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

Answer: Twigs