Assignment 3: Data Exploration

Maggie O'Shea, Section #1

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/maggieoshea/Desktop/Spring 2022/Data Analytics/Environmental_Data_Analytics_2022/Assignme
```

```
setwd(
  "/Users/maggieoshea/Desktop/Spring 2022/Data Analytics/Environmental_Data_Analytics_2022")
library(tidyverse)

#../ Did not work in accessing my Data/Raw folder, thus I wrote out the whole path.
neonics<-read.csv(
  "/Users/maggieoshea/Desktop/Spring 2022/Data Analytics/Environmental_Data_Analytics_2022/Data/Raw/ECO
  stringsAsFactors = TRUE)

litter<-read.csv(
  "/Users/maggieoshea/Desktop/Spring 2022/Data Analytics/Environmental_Data_Analytics_2022/Data/Raw/NEO
  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: One application of this work, understanding the ecotoxicology of neonicotinoids on insects, is that these pesticides can have harmful implications for bees. Bees are not only ecologically important but also economically important. Some of the compounds in these pesticides are toxic to bees and can prove to be a strong threat to the pollinator which has both ecological and socio-economic implications. **Source:** Decourtye, A. and Devillers, J. 2010. "Ecotoxicity of neonicotinoid insecticides to bees." *Advances in Experimental Medicine and Biology* 683: 85-95. doi: 10.1007/978-1-4419-6445-8_8.

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 can be an important dataset for a number of reasons. One that is specific to Colorado and the Western United States broadly is due to the outbreak of the mountain pine beetle. Woody Debris can offer insight as to how many trees were lost to an outbreak compared to trees lost in a stand without an outbreak. This was examined in the research by Klutsch et al. where they examined this to understand if the species associated with the outbreaks, lodgepole pine, will remain the dominant overstory tree after the beetle has moved through the forest. Understanding litter and woody debris in a forest can also be important to understand fuel loads prior to a wildfire season as litter and woody debris can act as fuel for these fires. **Source:** Klutsch, J., Negron, J., Costello, S., Rhoades, C., West, D., Popp, J., & Caissie, R. 2009. "Stand characteristics and downed woody debris accumulations associated with a mountain pine beetle (*Dendroctonus ponderosae* Hopkins) outbreak in Colorado." *Forest Ecology and Management* 258(5): 641-649. https://doi.org/10.1016/j.foreco.2009.04.034.

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 is sampled from elevated traps and woody debris is sampled from ground traps from NEON sites that contain woody vegetation greater than 2 meters tall. The litter sampling takes place in 20 40-meter-by-40-meter plots in sites with forested tower airsheds. Where there is low-statured vegetation over the tower airsheds, litter sampling is done in 4 40-meter-by-40-meter plots, and 26 20 meter by 20 meter plots. Ground traps, for woody debris, are sampled once a year whereas elevated trap sampling can vary depending on the vegetation at the site. * The finest resolution of temporal data is the range between the set data and the collect data, which is the day of the collection of the individual trap. This is called the "daysOfTrapping". * Each elevated trap (litter) is paired with a ground trap which makes for no more than 40 ground traps per site. * Traps can be placed in either a targeted or randomized way. This is determined based on vegetation. For example, if a site has greater than 50% aerial cover of woody vegetation greater than 2 meters tall, then the litter traps are random. If the site has less than 50% cover with patchy vegetation, then traps are placed only beneath qualifying vegetation.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
ncol(neonics)
```

```
## [1] 30
```

```
nrow(neonics)
```

```
## [1] 4623
```

Answer: The neonics dataframe has 30 columns and 4623 rows.

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 that are studied are population and mortality. These effects may be of particular interest because when examining the ecotoxicology of insects humans interact with or rely on the researchers may be concerned about the fatally harmful impacts of these insecticides. Not only may this be the case for species humans are concerned about, such as pollinators, but also for species that are considered pests it may be of interest to know if the insecticide is effective in reducing the population or increasing the mortality of the 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
```

Answer: The Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, Italian Honeybee are the six most studied species. All of these species are pollinators thus, may be studied more frequently because of the key role that pollinators play in the human food systems and economics.

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"
```

```
mode(neonics$Conc.1..Author.)
```

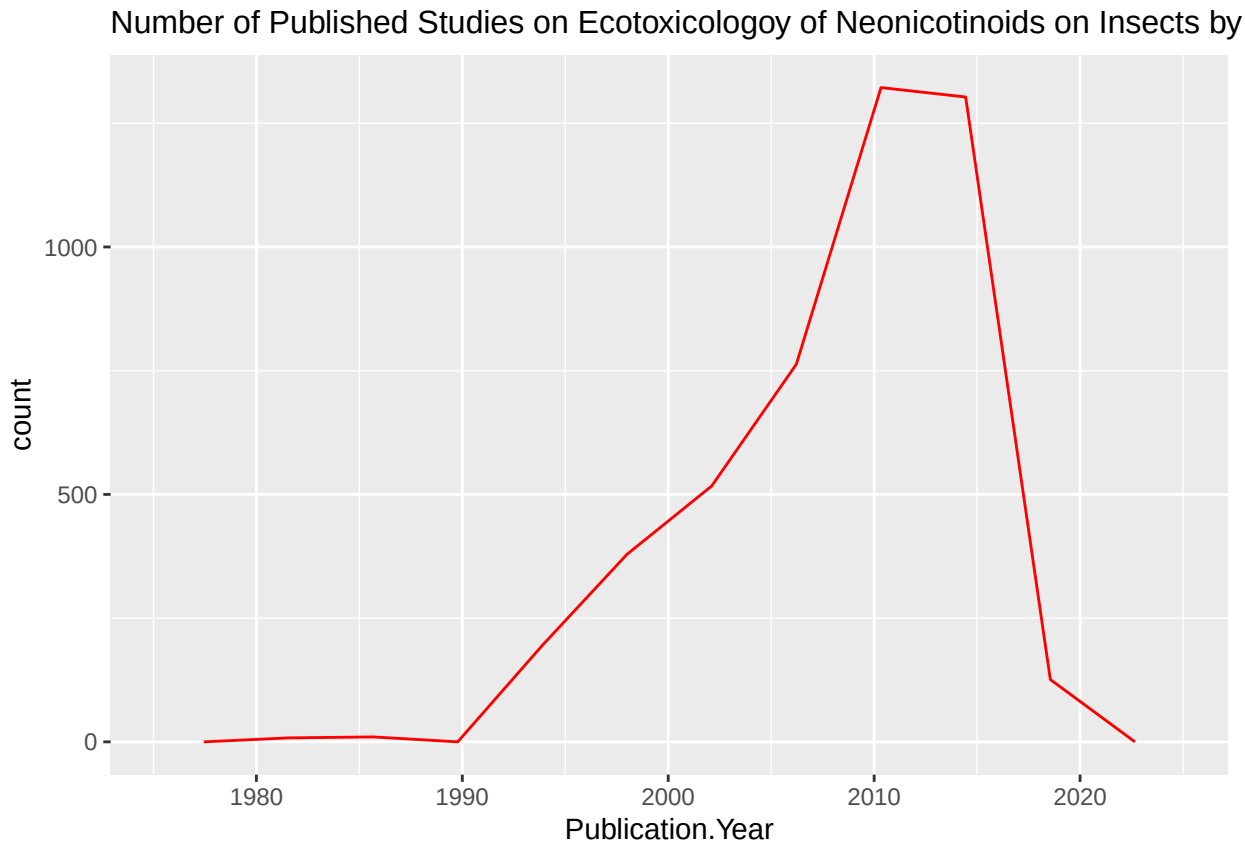
```
## [1] "numeric"
```

Answer: This column's class is "factor" which is a vector with numeric codes. This means that they are distinct from numeric because codes do not necessarily function as numeric values that can be added/subtracted/etc, but rather are values that represent something. Another example may be 0,1 indicating yes or no - these numbers represent something at the character level (yes, no) and thus do not function like a typical numeric vector. This may be because some of the values in Conc.1..Author are not numbers and are listed as NR and others have a / following the number, thus it cannot be understood as a solely numeric column.

Explore your data graphically (Neonics)

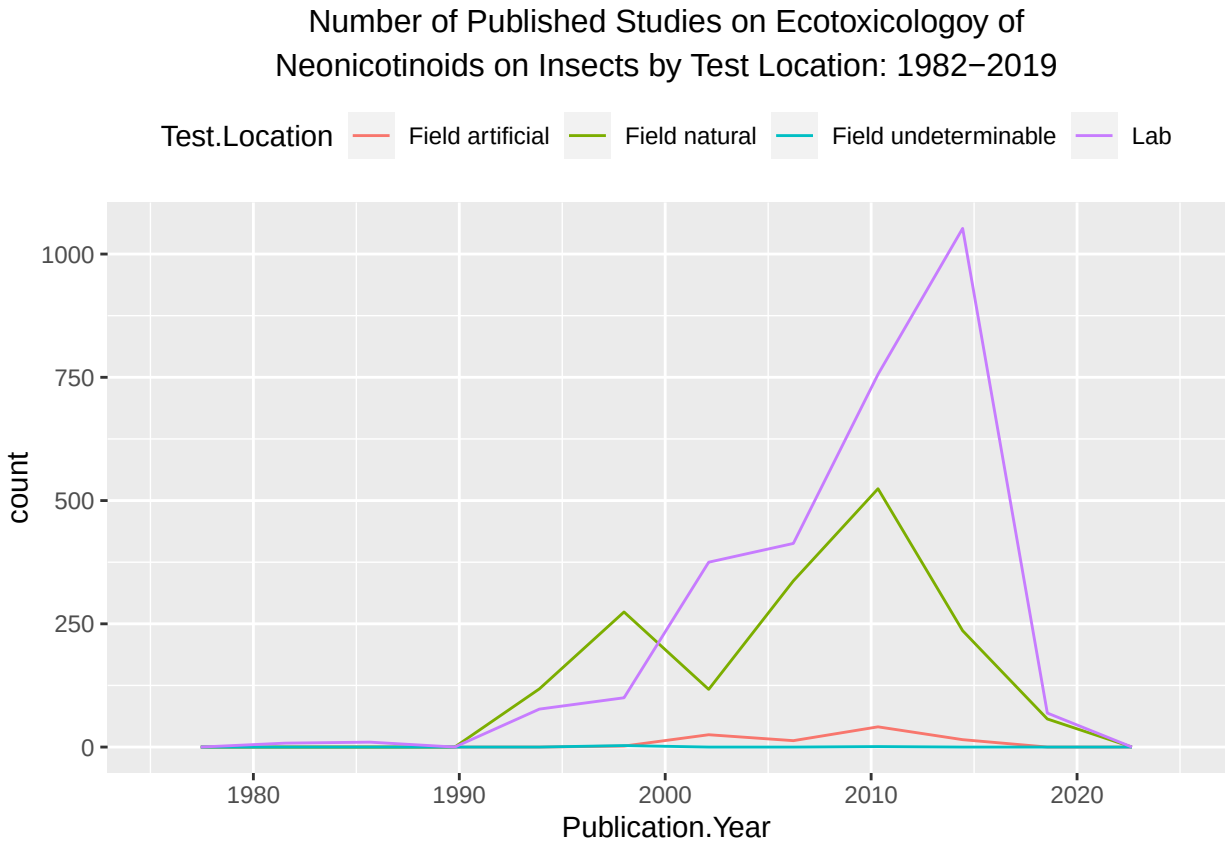
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 10, color="red") + labs(title="Number of Published St
    theme(plot.title = element_text(size=11.5))
```



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 = 10) + labs(title="Number of Pu
  theme(legend.position = "top")+
    theme(plot.title = element_text(size=11.5, hjust = 0.5), plot.subtitle = element_text(size=11.5, hj
```

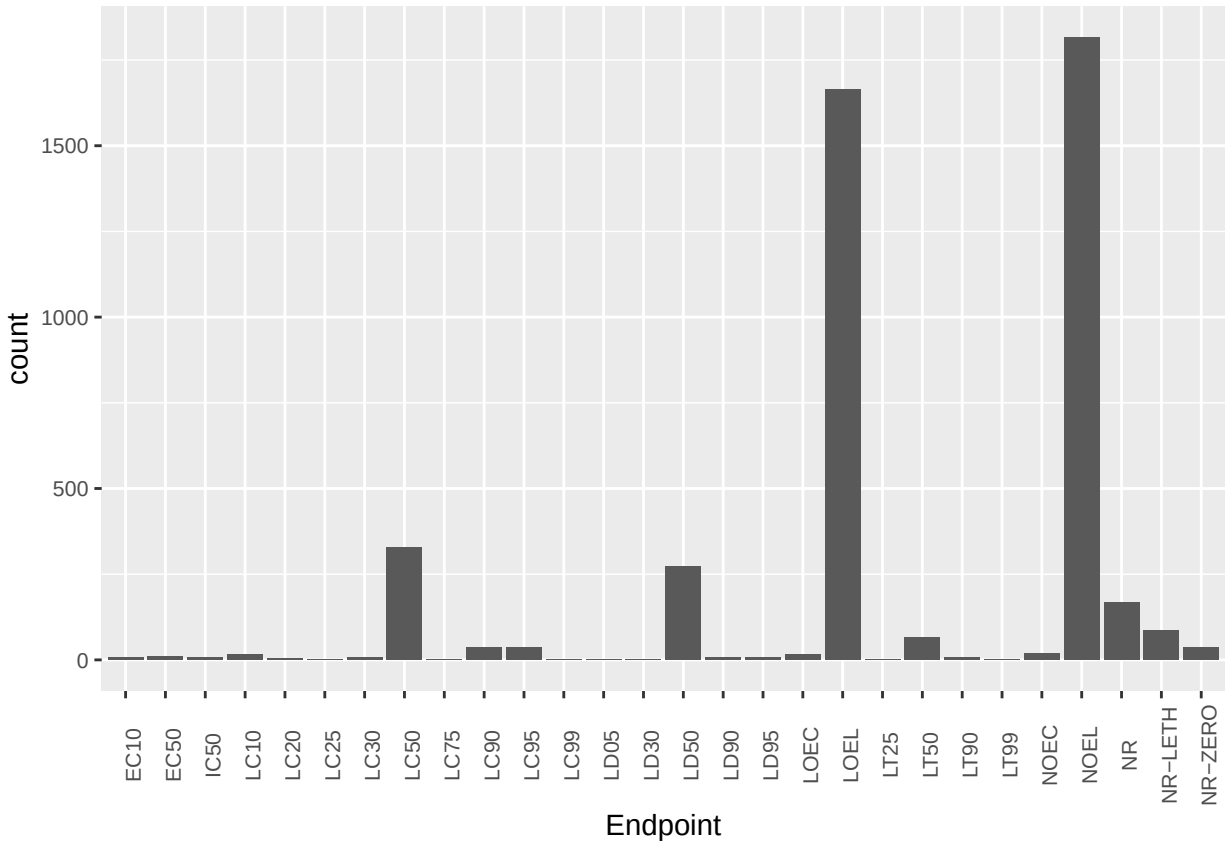


Interpret this graph. What are the most common test locations, and do they differ over time?

> **Answer:** The most commmon test locations are the "field natural" and lab locations. This graph shows that studies that use field artificial did peak in 2010, however are extremely rare compared to natural field and lab studies. Lab studies have increased significantly over time. Prior to 2005 natural field and lab studies had comparable number of studies, however by 2010 the number of lab studies far exceeded the number of natural field. For comparison, natural field studies appear to have peaked in 2010 at 500 studies while lab studies reached as high as 1000 studies around 2015. This could indicate that research is trending towards more lab studies, though further analysis may be needed to determine this.

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()+theme(axis.text = element_text(size = 8), axis.text.x = element_text(angle = 90))
```



Answer: The two most common endpoints are LOEL and NOEL. LOEL stands for the "Lowest-observable-effect-level" which indicates that the study had the lowest dose produce effects that were significantly different than the response from the controls. NOEL stands for "No-observable-effect-level" which indicates that the study had the highest dose that produced effects not have significantly different responses from that of the 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"
```

```
#collectDate is not a date, but a factor. The values are listed as full year-month-day.

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

class(litter$collectDate)
```

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

```
unique(litter$collectDate)
```

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

Answer: Litter was sampled on August 2nd and August 30th in 2018.

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
```

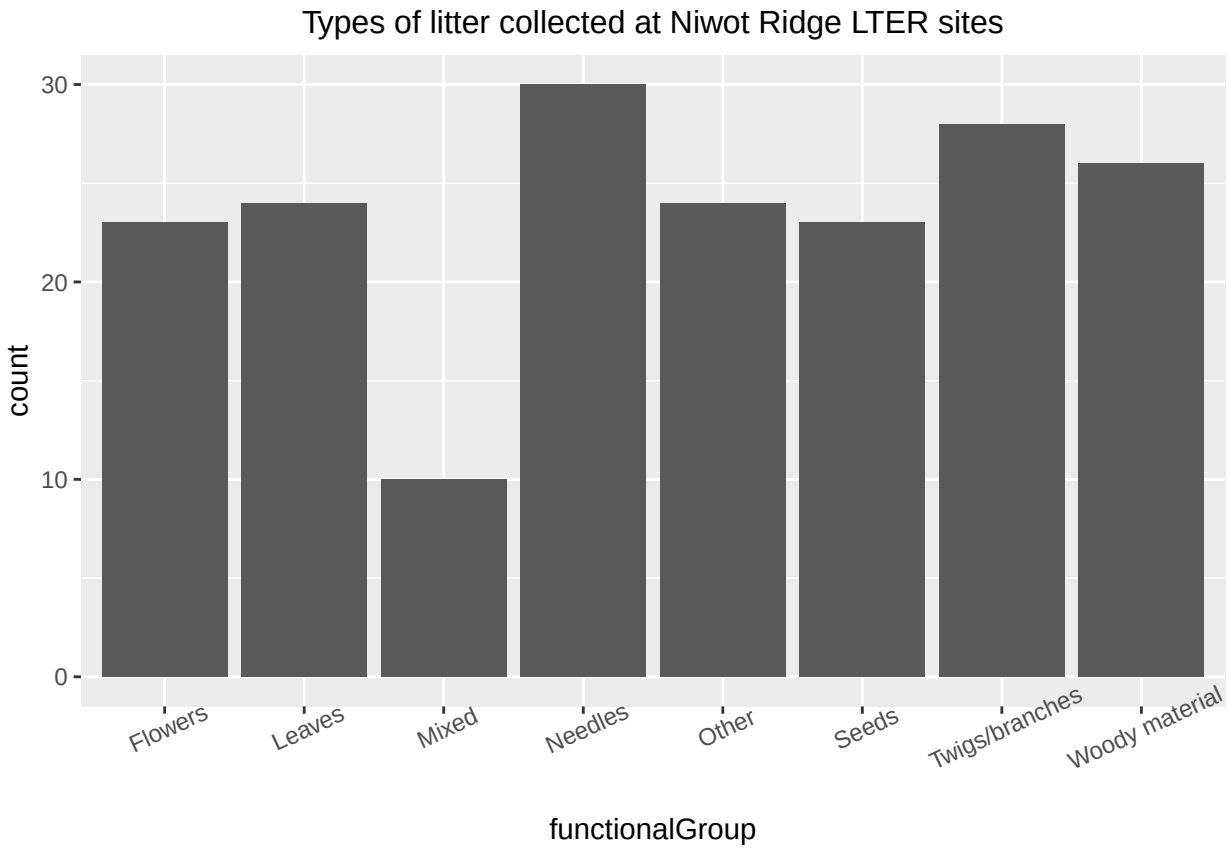
```
summary(litter$plotID)
```

```
## 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: When using summary with the code: summary(litter$collectDate) the information returned is the number of observations that have a certain value with each unique plot. For example, summary returned that 20 observations were collected at the plot with ID NIWO_040, and 19 and plot NIWO_041. In contrast, unique returns the values in a particular vector or dataframe that are unique. Rather than showing the number of observations that have this same observation it returns just the observations that appeared once excluding any duplicates. In this case, the unique command returned 12 unique plot IDs, indicating that 12 plots were sampled at Niwot Ridge.

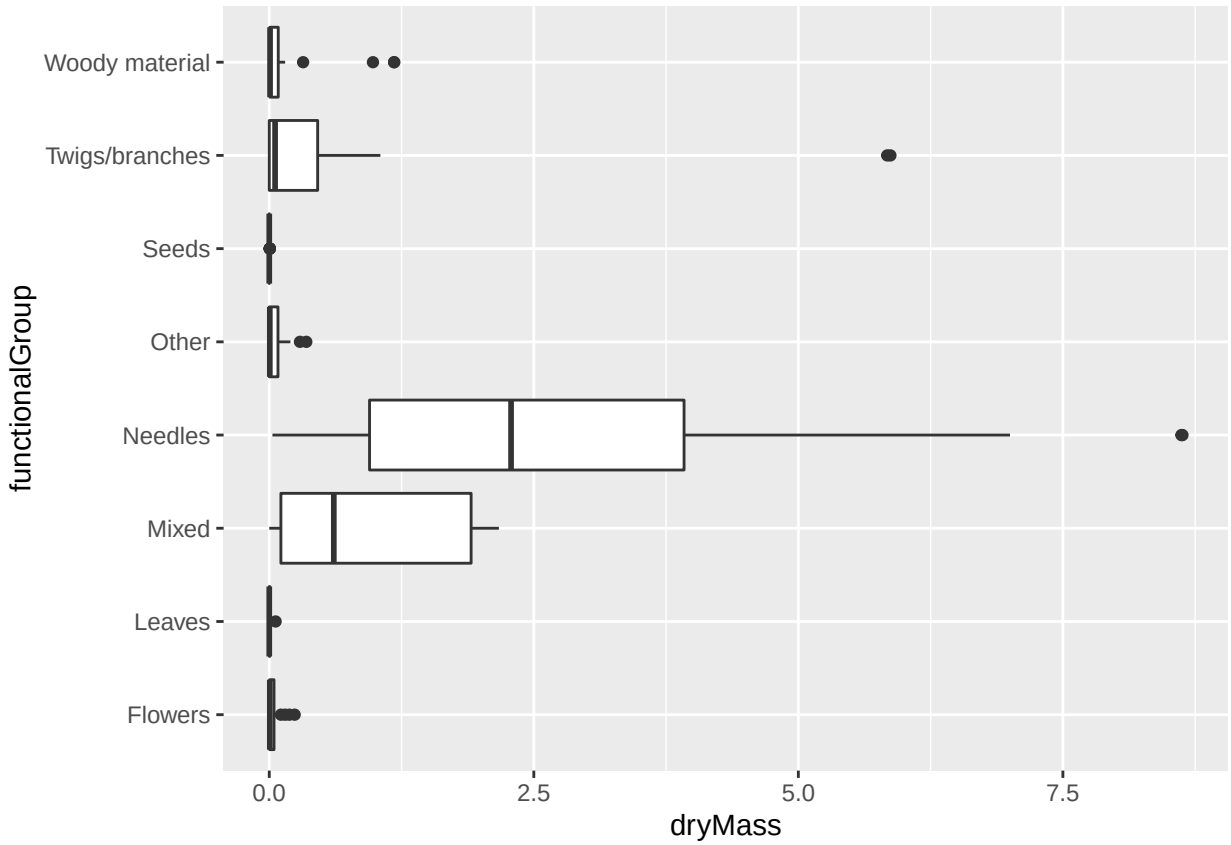
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(litter)+ aes(x = functionalGroup) +
geom_bar()+theme(axis.text = element_text(size = 9), axis.text.x = element_text(angle = 25), plot.title
```



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

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

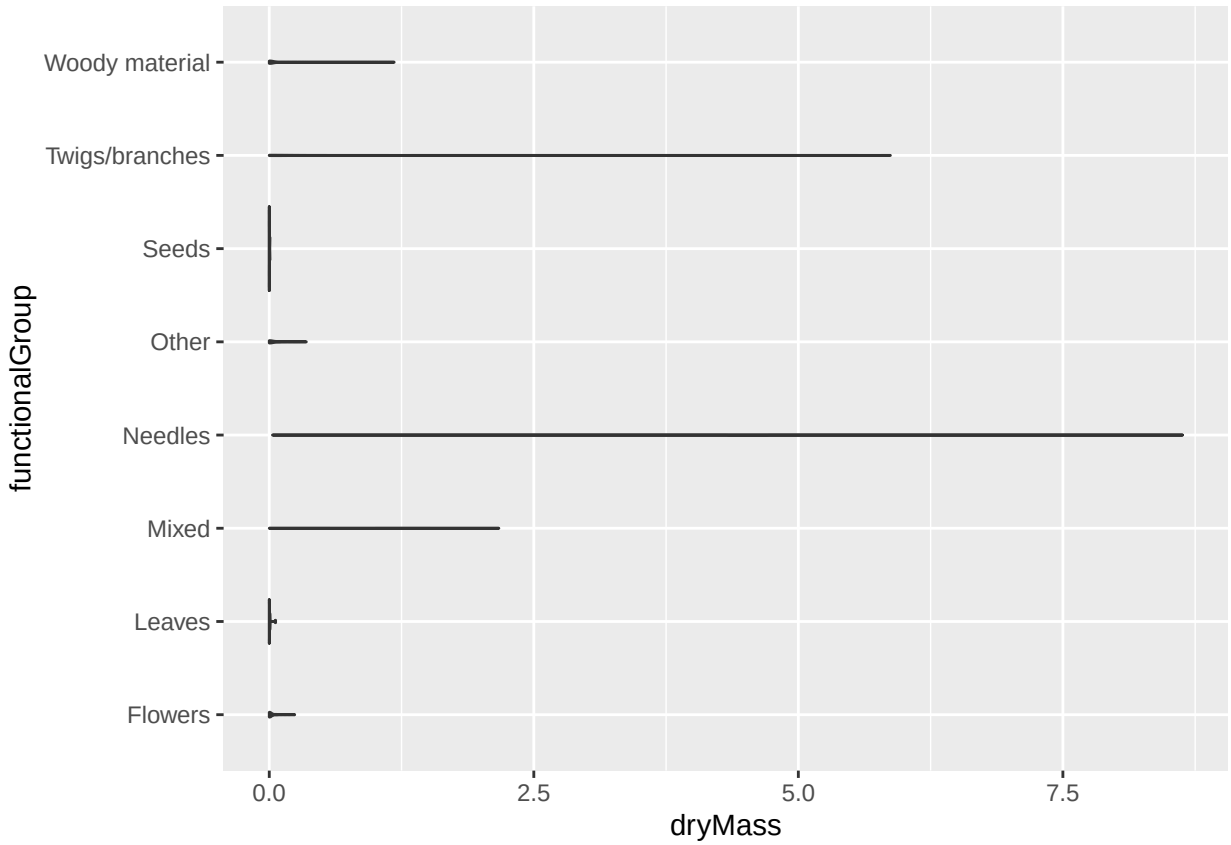



```
#violin shows how much data are in each distribution area
ggplot(litter) +
  geom_violin(aes(x = dryMass, y = functionalGroup),
              draw_quantiles = c(0.25, 0.5, 0.75))
```

```
## 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 a more effective visualization because the violin plot relies on observations having the same value for, in this case, DryMass which is not the case in this dataset. Violin plots help to show how much data are in each distribution area, and may appear wider where many observations have a particular value. However, because it appears that there is often only one or few observations with the same dryMass value, the violin plot just produces straight lines that are difficult to glean information from. In contrast, the box plot does not change its width according to the number of observations at a particular value, and thus shows more clearly key summary statistics such as the median, as well as the interquartile range and the maximum and minimum.

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

> **Answer:** Needles appear to have the highest biomass at sites.