Assignment 3: Data Exploration

Rachel Gordon, 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.**

```
setwd
```

```
## function (dir)
## .Internal(setwd(dir))
## <bytecode: 0x15821abf8>
## <environment: namespace:base>
```

```
("~/Documents/Duke Spring 2022/Environmental Data Analytics/Environmental_Data_Analytics_2022")
```

```
## [1] "~/Documents/Duke Spring 2022/Environmental Data Analytics/Environmental_Data_Analytics_2022"
```

```
getwd()
```

```
## [1] "/Users/rachelgordon/Documents/Duke Spring 2022/Environmental Data Analytics/Environmental_Data_A
```

```
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: It's important to study the ecotoxicology of neonicotinoids on insects because neonicotinoids can have determintal risks to bees and other pollinators, negatively affecting their immune systme and making them more susceptible to parasites and disease. Neonicotinoids are especially toxic to bees, as they become present in pollen and nectar when they are absorbed by the plant, posing risks to the bees and other pollinators when they feed on the plant. As neonicotinoids are important for managing pests that are resistant to other insecticides, it's important to understand how using them for agricultural purposes may have negative impacts on bee and other pollinator populations.

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: It is important to study litter and woody debris that falls to forest grounds as they play a major role in forest biodiversity, nutrient cycling, and habitats for terrestrial organisms. Specifically, the coarseness of the litter and woody debris has an impact on the forest ecosystem, as more coarse debris can provide nesting, foraging, and shelter services to the various organisms living in the forest.

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: * The litter and woody debris are collected from elevated and ground traps * Ground traps are sampled once per year, whereas the sampling frequency of elevated traps depends on the vegetation present at the site, with sampling methods being 1x every 2week or 1x every 1-2 sites for deciduous forest sites vs. evergreen sites, respectively * Litter and fine woody debris sampling is done at terrestrial NEON sites that have woody vegetation >2m tall and sampling only occurs in tower plots

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#Using the dim() function to determine how many rows and columns there are.
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?

```
#Using the summary() function to determine the most common effects studied.
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, Mortality, and Behavior. As this data is helping study how neonicotinoids may be impacting insects, population and mortality are important to study as they are both effects linked to death/the health of a population. As neonicotinoids are known to be toxic to pollinators, studying the mortality and population of them can provide insight into the health of the populations. Additionally, neonicotinoids impact the nervous system functions, so studying behavior is also important to see if neonicotinoid exposure can be linked to behavior changes in pollinator populations that may be driven by nervous system disruptions.

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.

```
#Using the summary() function to identify 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 Bees, Parasitic Wasps, Buff Tailed Bumblebees, Carniolan Honey Bees, Bumble Bees, and Italian Honeybees. All of these species are part of the bee or wasp family, which are all important for studying neonicotinoids because as mentioned earlier, neonicotinoids pose a major risk and can be toxic to pollinators, such as bees/wasps.

8. Concentrations are always a numeric value. What is the class of Conc.1..Author. in the dataset, and why is it not numeric?

```
#Finding class of Conc.1..Author
class(Neonics$Conc.1..Author)
```

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

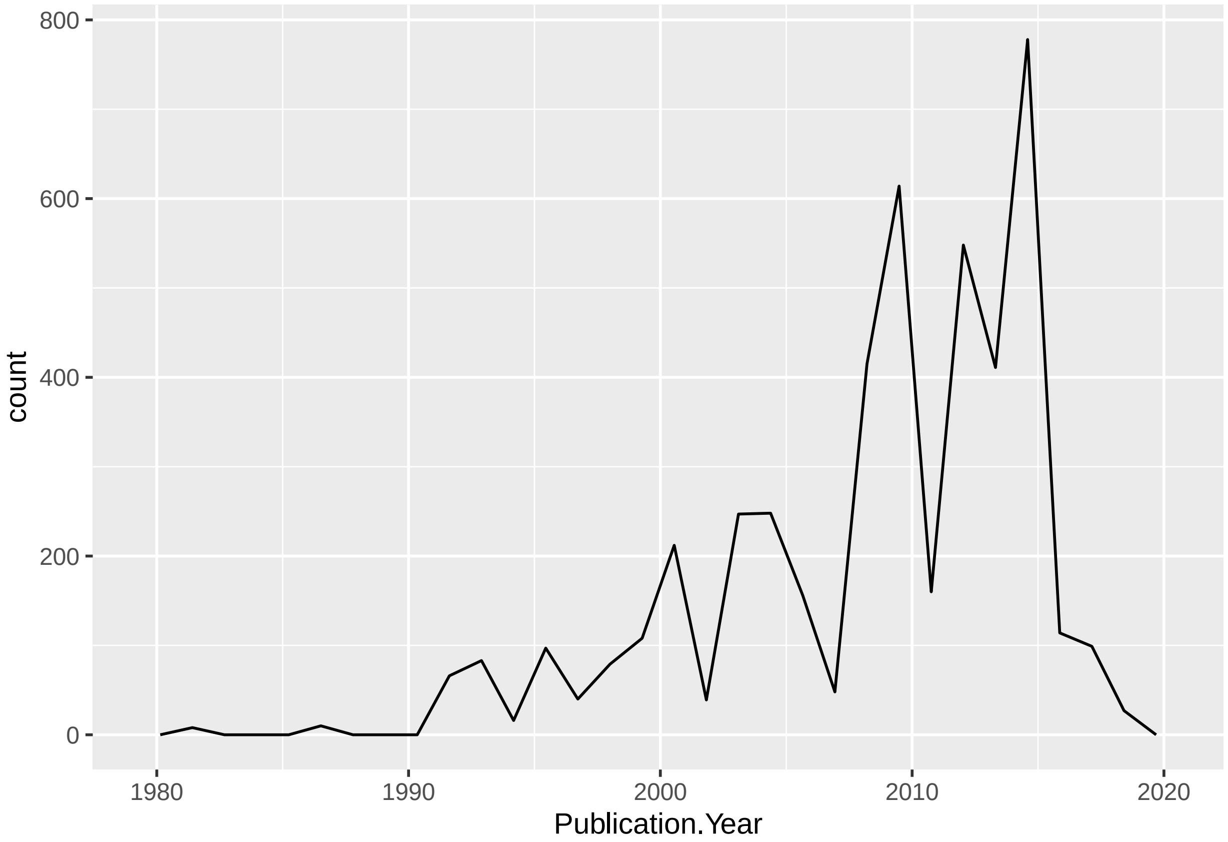
> Answer: The class of Conc.1..Author is ‘factor’. The class is not numeric because some of the entries in the data have ‘.’ or ‘/’ in them. The data will only be categorized as numeric by R if they’re all in numeric format and don’t contain any ‘.’,‘/’, or other symbols/characters that aren’t numbers. Since some of them do, R recognizes them as factors instead.

Explore your data graphically (Neonics)

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

```
#Using ggplot() and geom_freqpoly() to generate a plot of the # of studies by pub year
ggplot(Neonics)+
geom_freqpoly(aes(x=Publication.Year,))
```

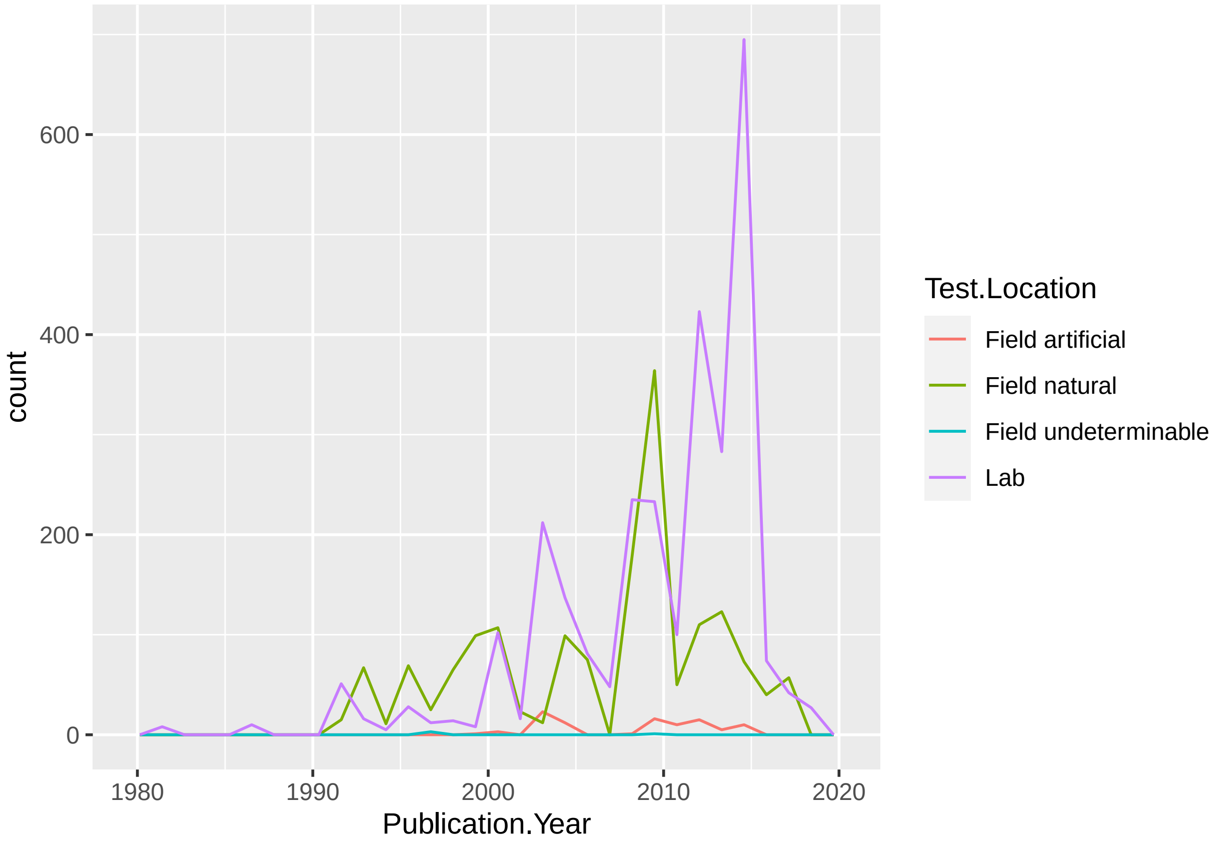
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
#Using ggplot() and geom_freqpoly() to add a color feature for different test locations
ggplot(Neonics)+
geom_freqpoly(aes(x=Publication.Year, color = Test.Location))
```

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

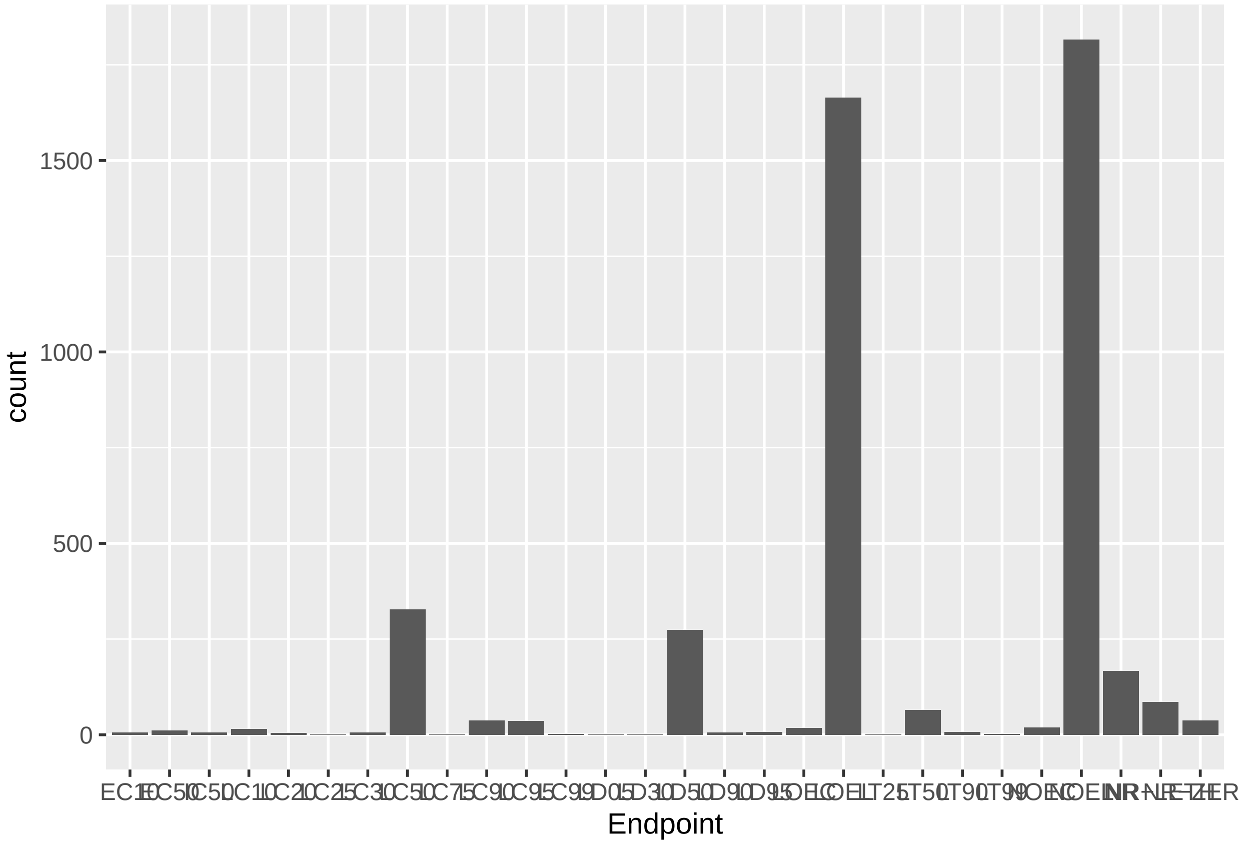


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are lab and field natural, with very few tests located at field undeterminable and field artificial. The test location trends have differed over time, with lab being most common from 1980-1990, then field natural being more common between 1990-2000, and eventually lab returning as the more common test location from 2000-2020.

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.

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



Answer: The two most common end points are NOEL and LOEL. The NOEL (terrestrial database usage) stands for no-observable-effect-level, meaning that the highest dose produces effects not significantly different from responses of constrols according to the author's report statistical test. The LOEL (terrestrial database usage) stands for lowest-observable-effect-level, meaning that the lowest dose produces effects that were significantly different from responses 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.

```
#Determining the class of collectDate
class(Litter$collectDate)
```

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

```
#Changing from Factor class to Date class
Litter$collectDate <- as.Date(Litter$collectDate)
#Confirming class of collectDate is now date
class(Litter$collectDate)
```

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

```
#Using 'unique' to determine which dates litter was sampled in August 2018.
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`?

```r
#Using 'unique' function to determine how many different types of plots there are by plotID
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
```

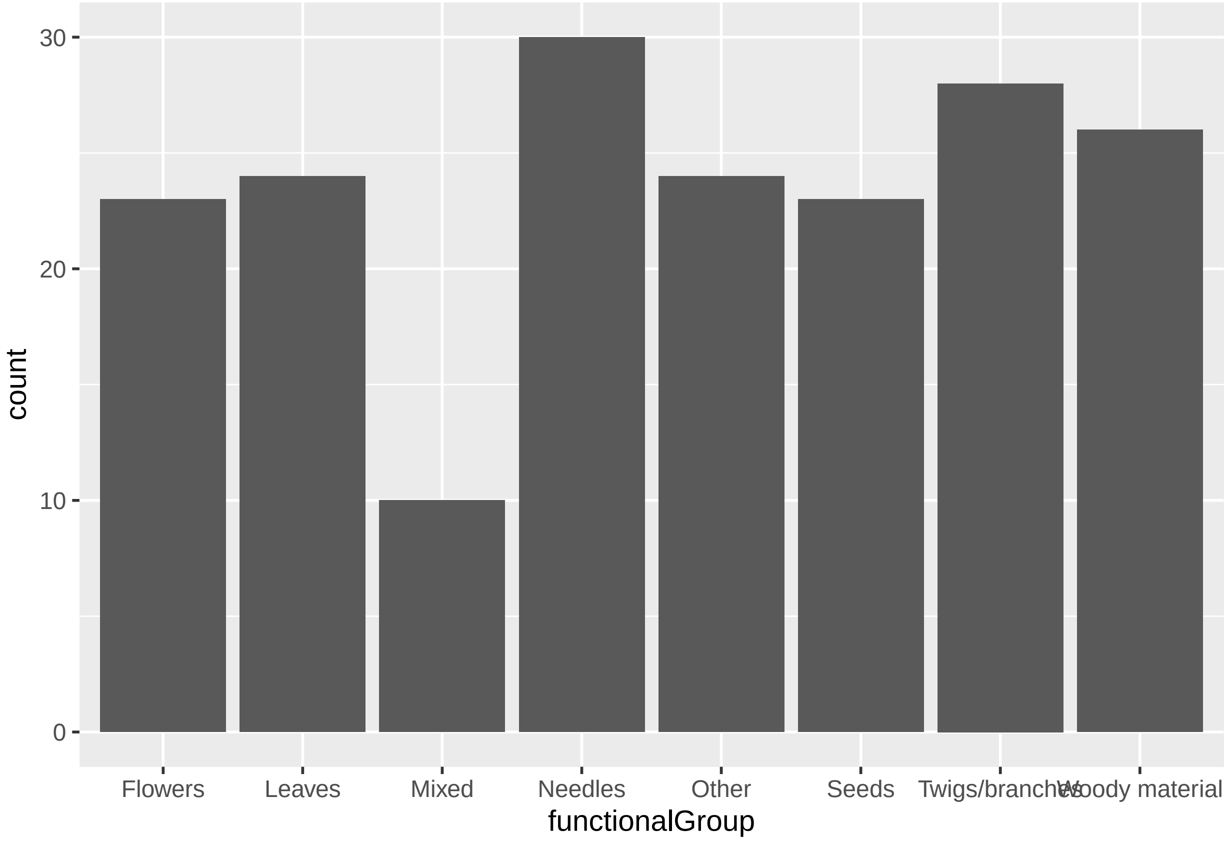
```r
#Using 'summary' function to determine the difference between 'unique' and 'summary'
summary(Litter$plotID)
```

```
## 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 were 12 plots sampled at Niwot Ridge. The information provided from ‘unique’ gives you an overview of how many unique levels there are for one specific column. The ‘summary’ function differs, as it provides a count of how many entries/occurrences of each of these unique levels there are in the dataset.

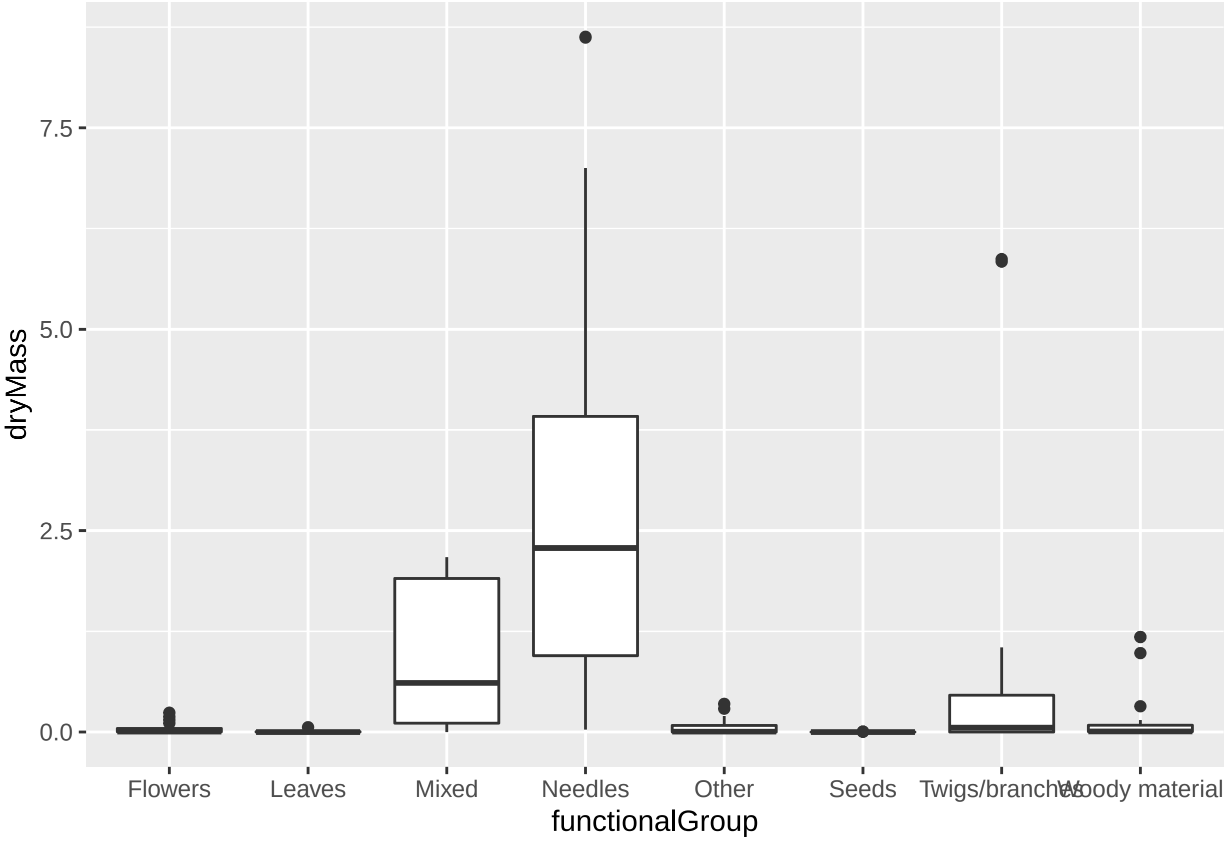
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
#Creating a bar graph of functionalGroup counts
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.

```
#Creating a boxplot of dryMass by functionalGroup
ggplot(Litter)+
  geom_boxplot(aes(x=functionalGroup, y=dryMass))
```

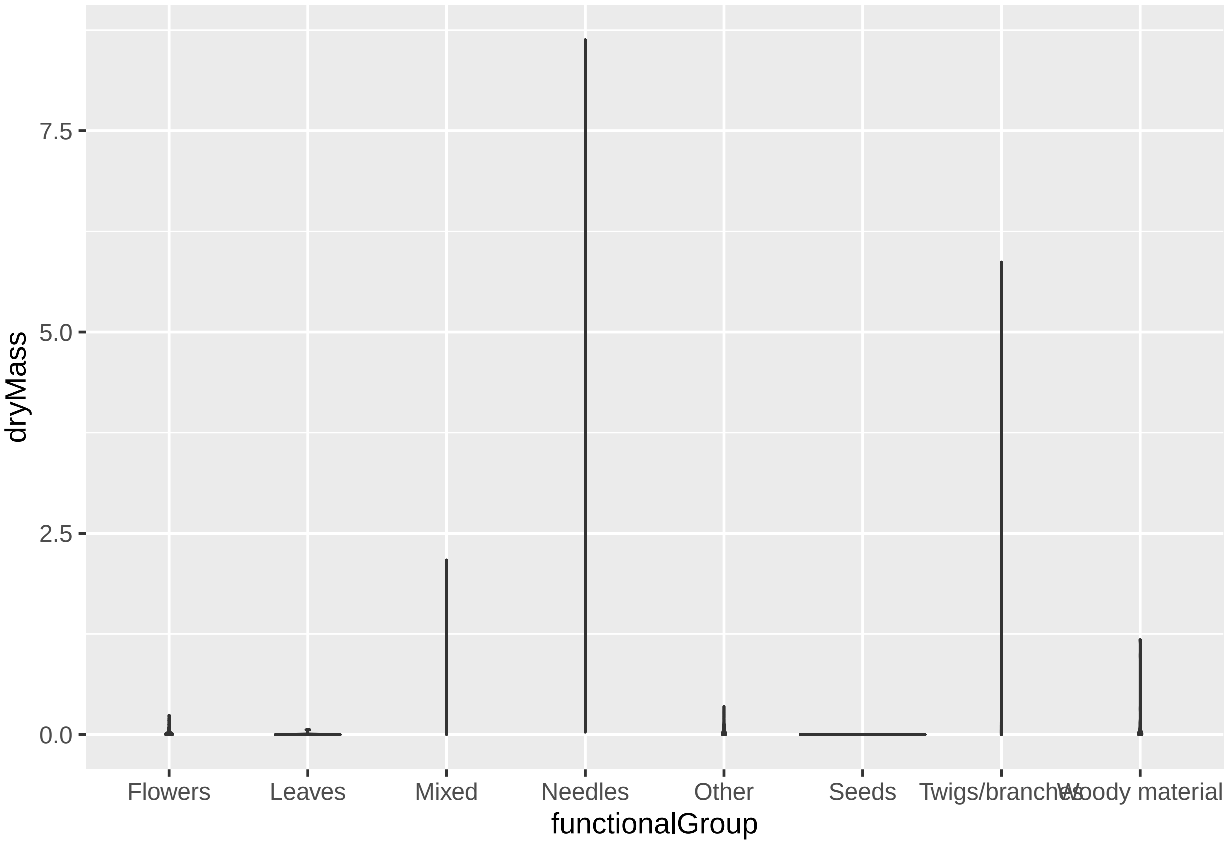


```
#Creating a violin plot of dryMass by functionalGroup
ggplot(Litter)+
  geom_violin(aes(x=functionalGroup, y=dryMass), draw_quantiles = c(0.25, 0.5, 0.75))
```

```
## 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 a more effective visualization option because we only want the summary statistics of the data, and the boxplot is effective at visualizing the summary points of the dryMass data by functionalGroup. The violin plot is not as effective in this case, as it shows the entire distribution of the data, which is not as useful to us in this case.

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

> Answer: Based on the plots, it is claer that 'Needles' has the highest biomass at these sites, followed by the 'Mixed' functional group.