Assignment 3: Data Exploration

Jess Garcia

OVERVIEW

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

Directions

1. Change "Student Name" on line 3 (above) with your name.
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., "Salk_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.

```
#Confirm working directory
getwd()
```

```
## [1] "C:/Users/93jes/Documents/ENV872/Environmental_Data_Analytics_2021"
```

```
#Correct working directory to main folder
setwd("~/ENV872/Environmental_Data_Analytics_2021")
getwd()
```

```
## [1] "C:/Users/93jes/Documents/ENV872/Environmental_Data_Analytics_2021"
```

```
#Install and load packages
#install.packages("tidyverse")

library(tidyverse)

#Upload & rename datasets
Neonics.data <-read.csv("~/ENV872/Environmental_Data_Analytics_2021/Data/Raw/ECOTOX_Neonicotinoids_Inse

Litter.data <-read.csv("~/ENV872/Environmental_Data_Analytics_2021/Data/Raw/NEON_NIWO_Litter_massdata_20
```

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: We might be interested in the ecotoxicology of neonicotinoids on insects, in particular to see the effects on pollinators such as bees. Even though they are a class of insecticides designed for the purpose of killing insects, there might be unintended consequences on the greater ecosystem as as result of these insecticides.

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: According to Science Direct, there is a lot of value to studying litter and woody debris in forests. It can impact stream channels through sediment retention, altering the water velocity, and changing the morphology of the streams and banks. Litter and wood are important sources of biomass in terrestrial and aquatic ecosystems especially as a source of energy and nutrients for microorganisms and detritivores (species that consume decaying materials). These materials can be an indicator of the health of an ecosystem, for example by studying them before and after fires or infestations.

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 collected from elevated PVC traps, each is a 0.5m2 square with mesh 'basket' elevated ~80cm above the ground. The fine woody debris is collected from ground traps that are 3 m x 0.5 m rectangular areas.* Spatial sampling: Sampling only takes place at sites where woody vegetation is greater than 2m tall. In areas where there is more than 50% aerial cover of woody vegetation, the litter trap placement is randomized. However, it sites with less than 50% woody vegetation the trap placement is targeted to areas beneath qualifying vegetation. *Temportal samping: The ground traps are sampled once per year.The elevated traps are sampled once every two weeks in deciduous forest sites with potential discontinuation during winter months, and sampled once every one or two months at evergreen sites.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#Obtain dimensions of dataset
dim(Neonics.data)
```

```
## [1] 4623   30
```

```
#There are 4,623 rows and 30 columns of data
```

6. Using the summary function on the "Effects" column, determine the most common effects that are studied. Why might these effects specifically be of interest?

```
#View data
View(Neonics.data)

#Get summary data of the "Effect" column
summary(Neonics.data$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 mortality and population. Mortality is critical because it's looking at where the cause of death is a direct result of the chemical. Population is an important factor because it is looking at whether the species in a given area are being impacted.

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.

```
#Get summary data on species in data set
summary(Neonics.data$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 data set are honey bees, parasitic wasps, buff tailed bummblebees, carniolan honey bees, bumble bees, and italian honeybees. These species might be of particular interest because they are prolific pollinators. Some of the other species studied are pollinators as wells, but bees are generally viewed as specialized for pollination and are the best at it. It is important for people to study insecticide impacts on pollinators, and particularly bees, because pollination is critical to human food production.

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 of the Conc.1..Author. datum
class(Neonics.data$Conc.1..Author.)
```

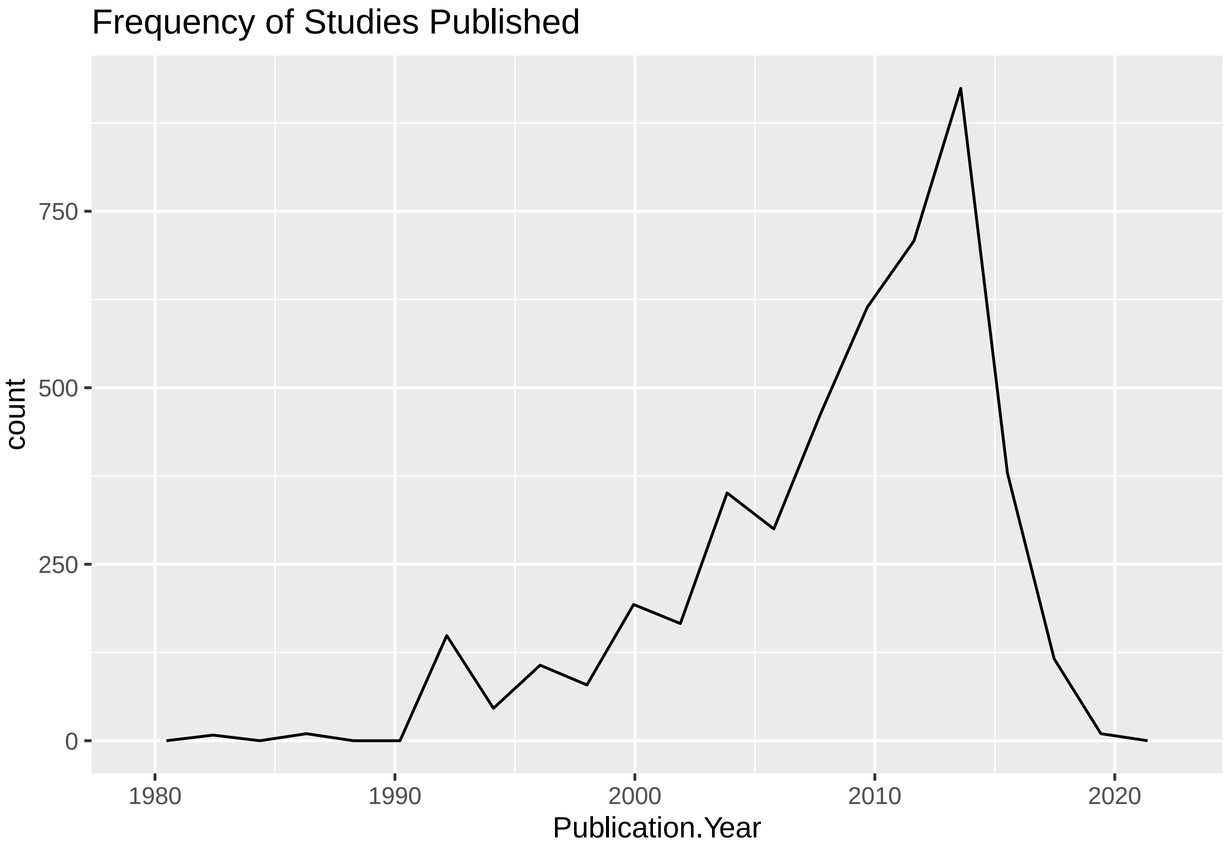
```
## [1] "factor"
```

Answer: The class of Conc.1..Author.is character. It is not numeric because some of the values are missing the denominator, instead a "/" was left, meaning that the concentration could not be calculated. *I could not find in the ECOTOX_CodeAppendix.pdf information on the column of interest. I found information about the concentration types but not the calculation of the concentrations, so I made an assumption about incomplete value calculations.

Explore your data graphically (Neonics)

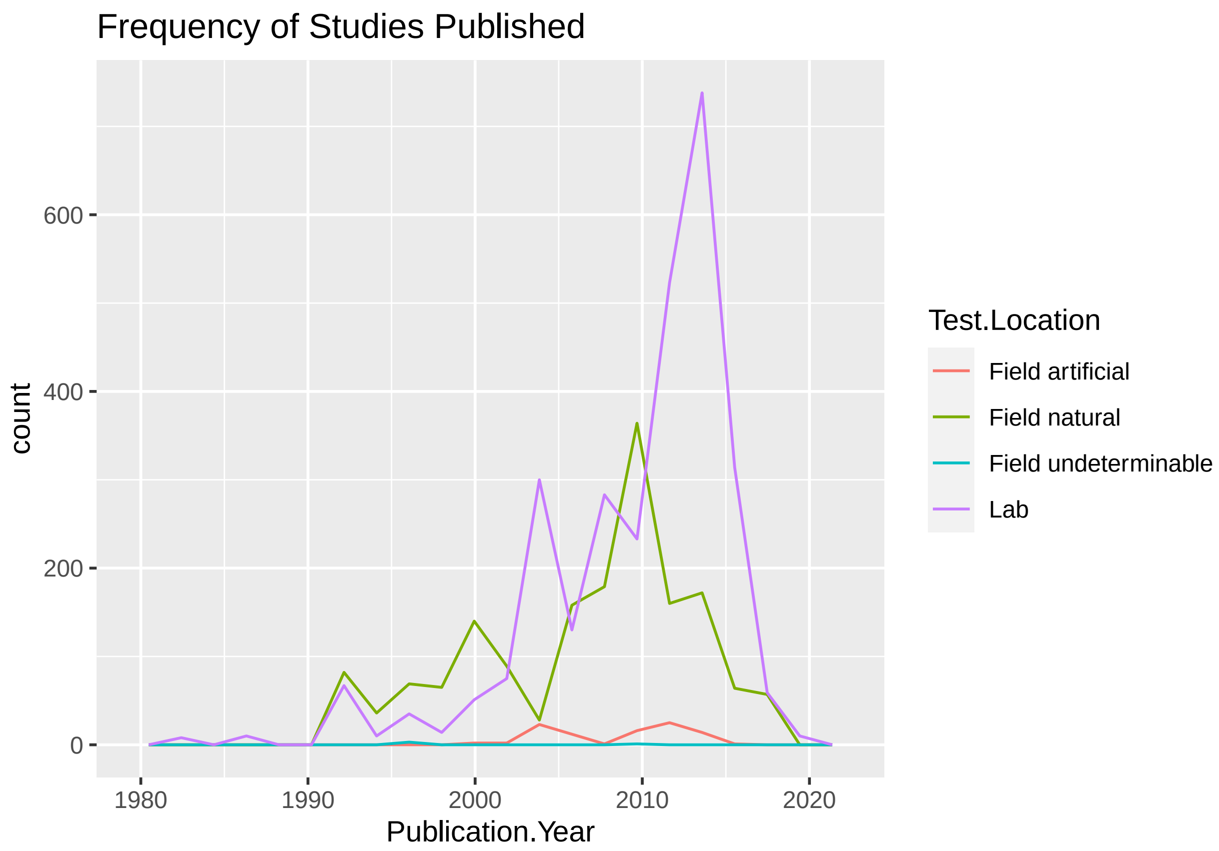
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
#Generating a plot of the number of studies conducted by publication year
ggplot(Neonics.data)+
geom_freqpoly(aes(x=Publication.Year), bins=20)+
  ggtitle("Frequency of Studies Published")
```



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

```
#Sorting publication year frequency by color based on Test.Location
ggplot(Neonics.data)+
geom_freqpoly(aes(x=Publication.Year, color = Test.Location), bins=20)+
  ggtitle("Frequency of Studies Published")
```

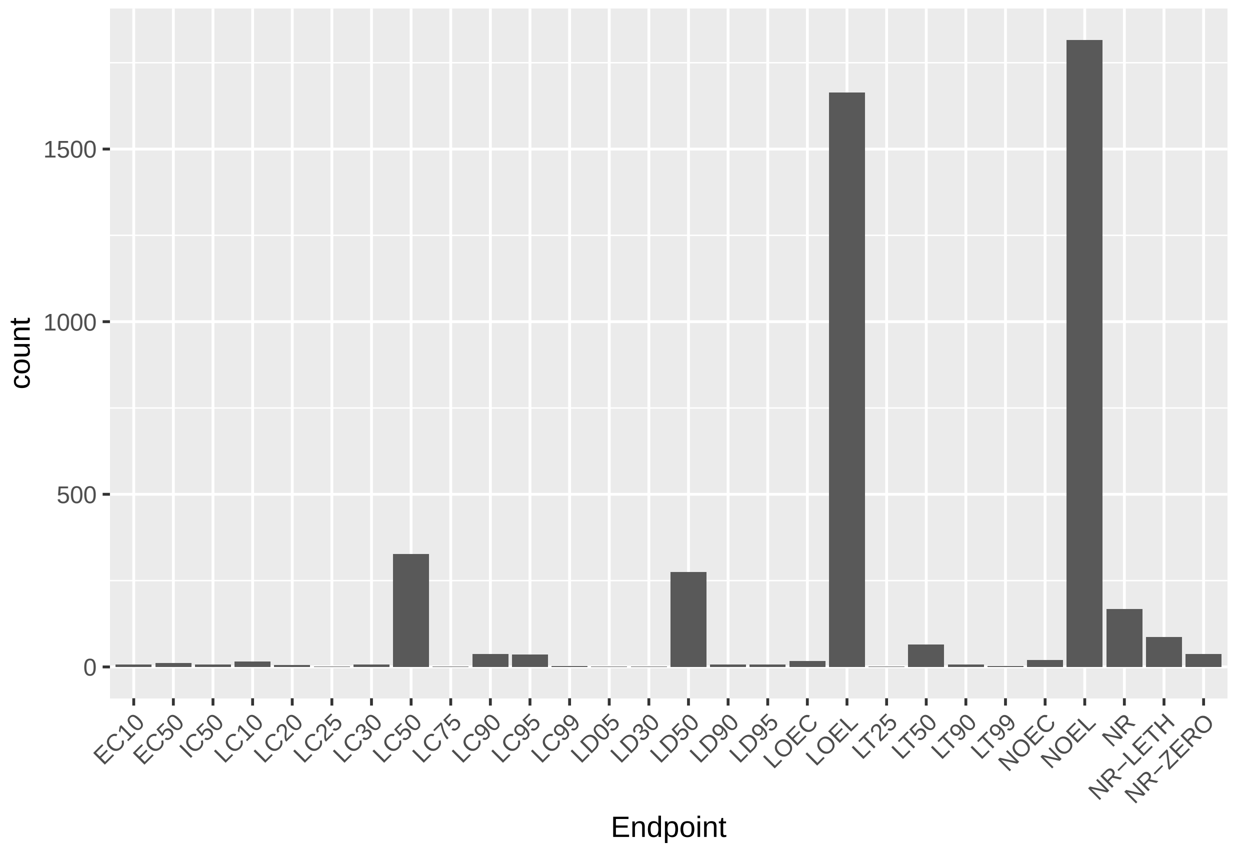


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are in the lab. However, before the year 2000 it was more common for test locations to be natural in the field. There are some artifical field test locations as well but not many.

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.

```
#Create bar graph of Endpoint counts
ggplot(Neonics.data)+geom_bar(aes(x=Endpoint))+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
  #This changes the angle of the x-axis labels to make them more legible
```

Answer: The two most common end points are LOEL and NOEL, both of which are used for terrestrial. LOEL is defined as the, “Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC)”. NOEL is defined as the, “No-observable-effect-level: highest dose (concentration) producing effects not significantly different from responses of controls according to author’s reported statistical test (NOEAL/NOEC)”.

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.

```
#View data
View(Litter.data)

#Determine the class of collectDate
class(Litter.data$collectDate)
```

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

```
#The class was imported as a factor variable, not a date variable

#Convert to date variable
Litter.data$collectDate <-as.Date(Litter.data$collectDate, formal = "%Y-%m-%d")

#Confirm new class of collectDate
```

```
class(Litter.data$collectDate)
```

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

```
  #Class confirmed as date

#Use unique function to determine which dates litter was sampled in August 2018
unique(Litter.data$collectDate)
```

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

```
#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`?

```
#Determining how many plots were sampled at Niwot Ridge
unique(Litter.data$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
```

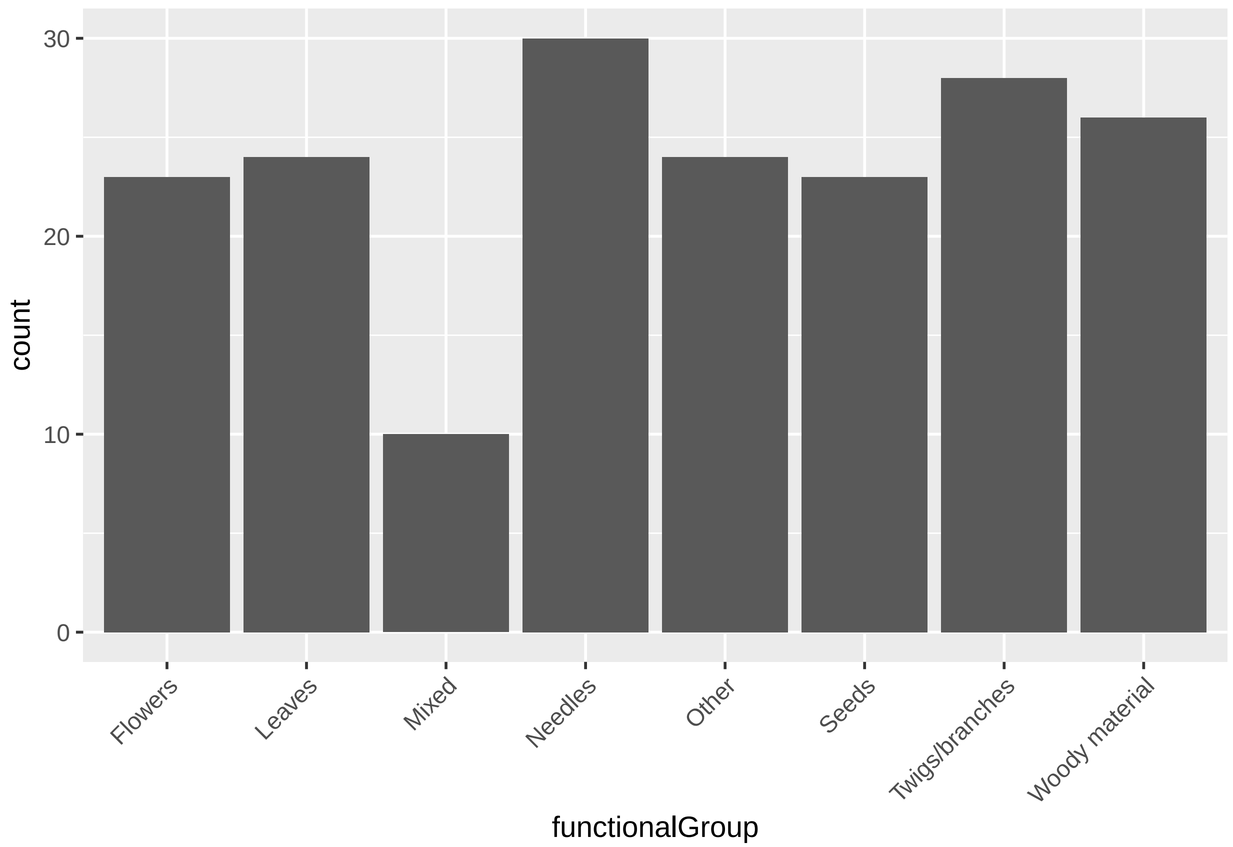
```
summary(Litter.data$plotID)
```

```
## 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: 12 plots were sampled at Niwot Ridge. The unique function just gives you the unique variables, so it will only tell you that there are unique variables. Whereas the summary function gives you the unique variables as well as how many times each of those variables is present in the dataset.

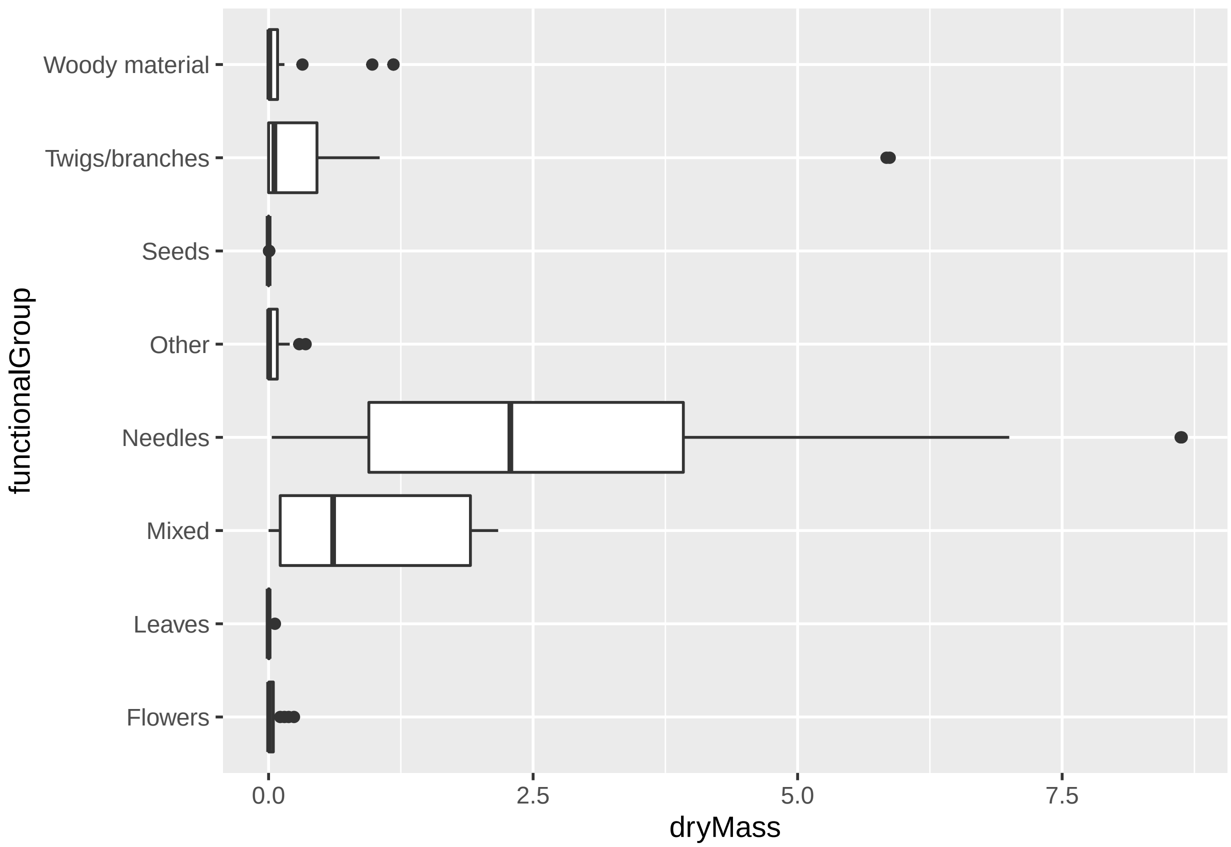
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#Create bar graph of functionalGroup counts
ggplot(Litter.data)+
  geom_bar(aes(x=functionalGroup))+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

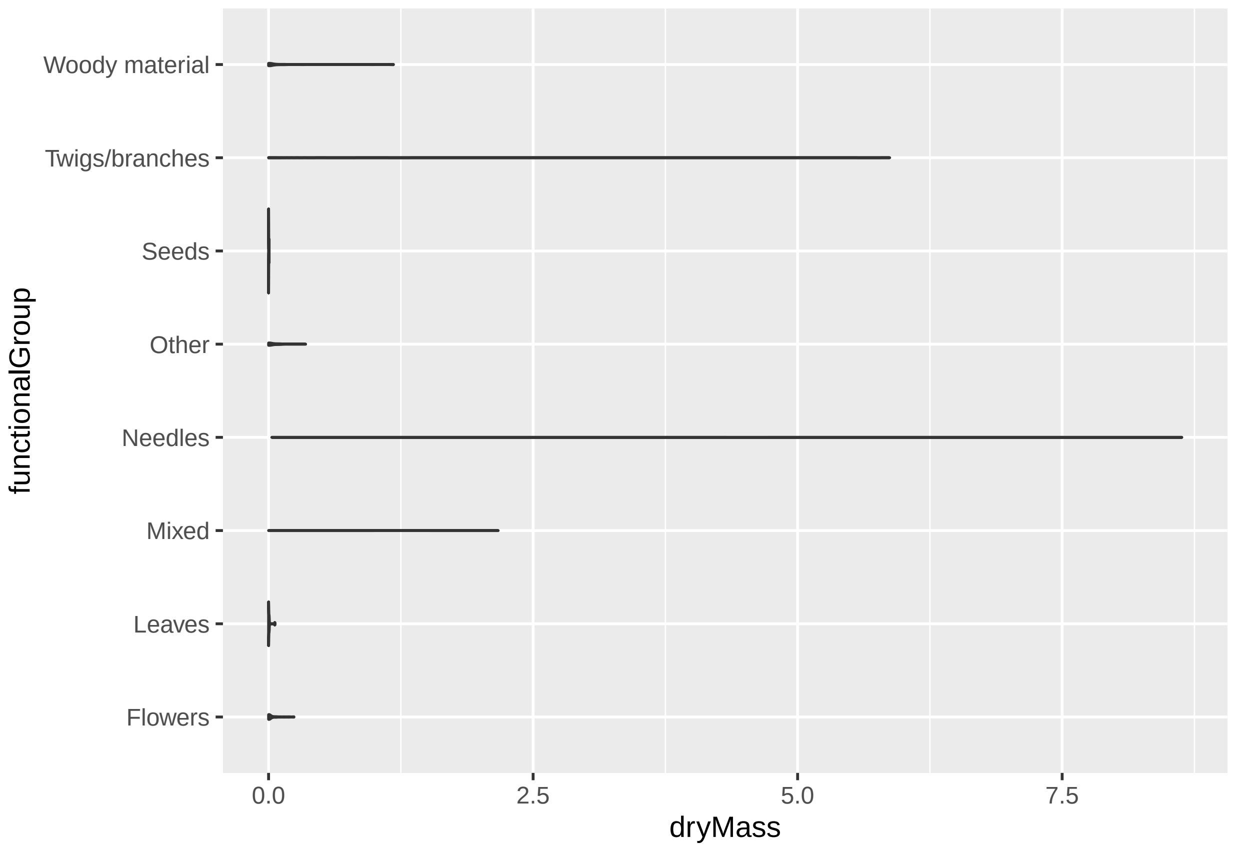


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#Create boxplot of dryMass by functionalGroup
ggplot(Litter.data)+
  geom_boxplot((aes(x=dryMass, y=functionalGroup)))
```



```
#Create violin plot of dryMass by functional group
ggplot(Litter.data)+
  geom_violin((aes(x=dryMass, y=functionalGroup)))
```



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

> Answer: I think the boxplot is a more effective vizualization option than the violin plot in this case because the data is continuous (decimals) rather than discrete so you aren't going to see much density at points.And there aren't thousands of data points, there are less than 200 collections (rows), so that means again you're not going to have a ton of repeat points which would expand out the violin plot. Instead you just get the points mostly along a line.

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

> Answer: The needles tend to have the highest biomass/dry mass at these sites, meaning the needles have the largest median. Mixed litter tends to have the next highest mass at these sites. The twigs might throw you off because they have a few outliers that have more mass, but the boxplot shows that is not the norm.