Assignment 3: Data Exploration

Taro Katayama, 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/tarokatayama/Desktop/Duke_Semester_2/Environmental_data_analytics/R_Projects/Environmenta
```

```
setwd("/Users/tarokatayama/Desktop/Duke_Semester_2/Environmental_data_analytics/R_Projects/Environmental
getwd()
```

```
## [1] "/Users/tarokatayama/Desktop/Duke_Semester_2/Environmental_data_analytics/R_Projects/Environmenta
```

```
library(tidyverse)
library(ggplot2)
Neonics<- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = TRUE)
view(Neonics)
Litter<- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv", stringsAsFactors = TRUE)
View(Litter)
```

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: We might be interested in the ecotoxicology of neonicotinoids for a variety of reasons. 1) we might be interested in the effects of neonicotinoids on human health. 2) we might be interested in the effects of neonicotinoids on wildlife, specifically non-target animals. 3) we might be concerned with the effectiveness of neonicotinoids on target species.

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: We might be interested in studying litter and woody debris that falls to the ground because they may indicate which species of insect lives within the leaf litter. For example, the spiders may like a certain type of leaf litter, and are found most ofetn there. It will be easier to identify the ideal habitat for these insect species, and apply the insecticide appropriately.

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 and woody debris was sampled at terrestrial NEON sites that contain woody vegetation that is greater than 2m tall.* In sites with forested tower airsheds, the samples took place in 20 40mx40m plots. *In sites with low-saturated vegetation over the tower airshed, the samples take place in 40mx40m tower plots + 26 20mx20m plots.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

   Answer: 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: population (1803 cases), followed by Mortality (1493), and Behavior (360) is the most common effects. These effects are of interest because they give you a sense of how effective the insecticide is to an assortment of insects.

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: Honey Bee (667), Parasitic Wasp (285), Buff Tailed Bumblebee (183), Carniolan Honey Bee (152), Bumble Bee (140) were the most commonly studied speices in the dataset. They are all bee/wasp species. They might "bee" of interest because bees are very important species that are responsible for a number of ecological services such as pollination and honey production. When developing insecticides, they may want to create one that does not have adverse effects on the important bee species.

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"
```

```
help("factor")
class(Neonics$Test.Location)
```

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

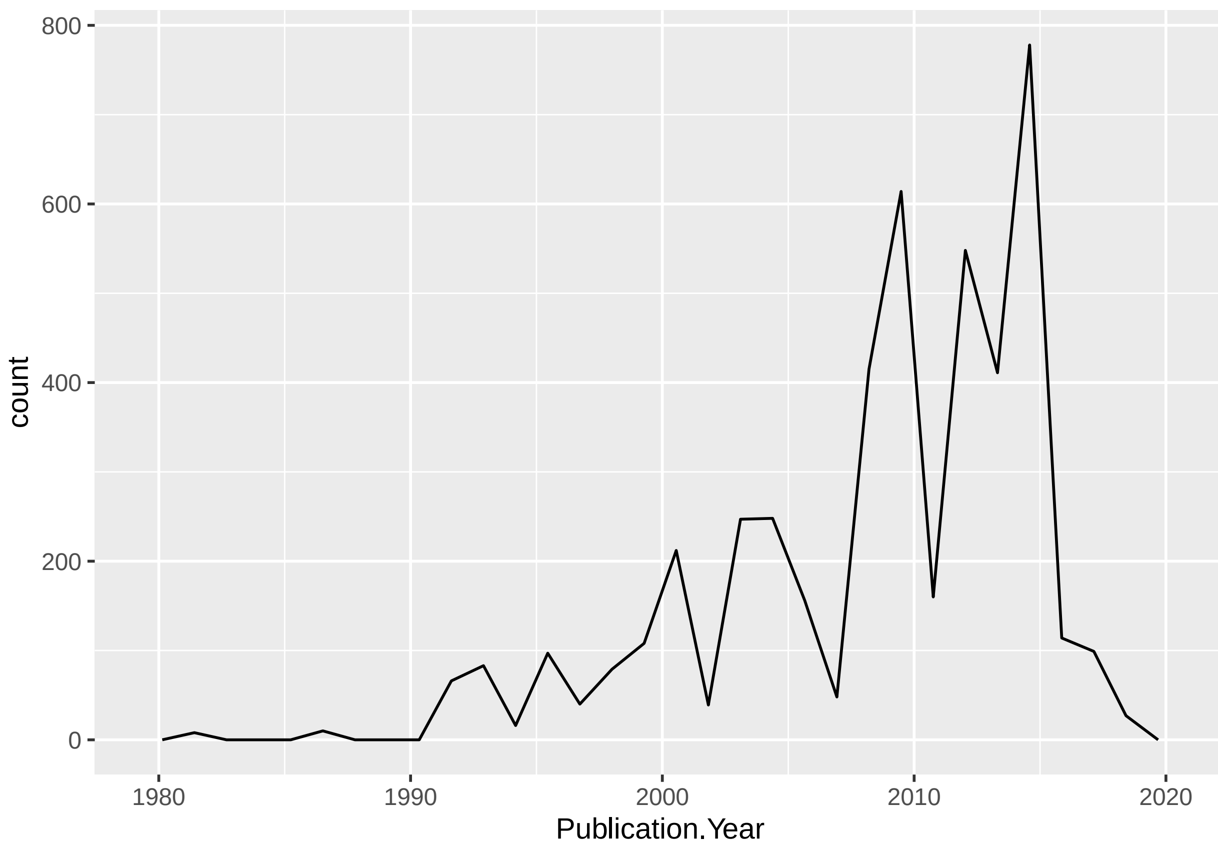
> Answer: The class is a factor. It is not numeric because the class Conc.1..Author is categorical data that has a limited number of different values. Categorical data is most appropriately classed as a factor.

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))
```

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

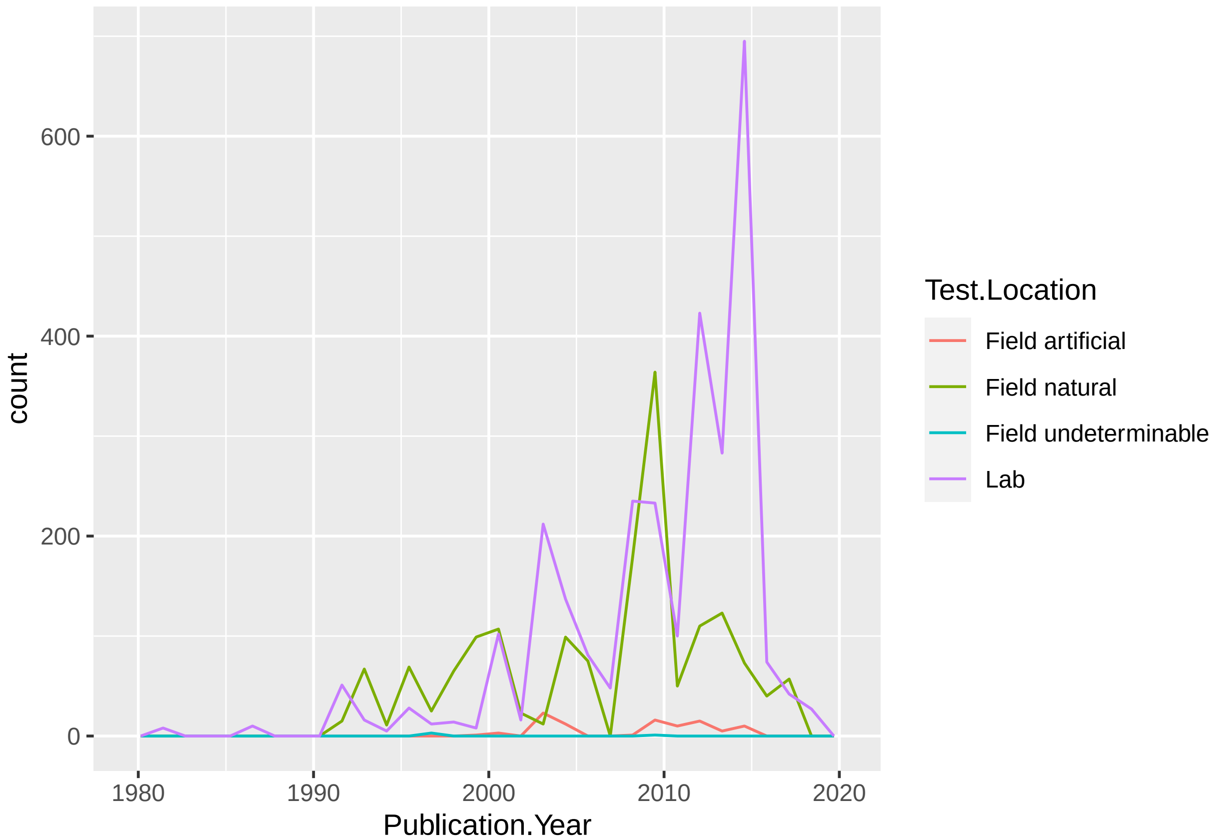


10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
ggplot(Neonics) +
  geom_freqpoly(aes(x=Publication.Year, color=Test.Location), bin=50)
```

```
## Warning: Ignoring unknown parameters: bin

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

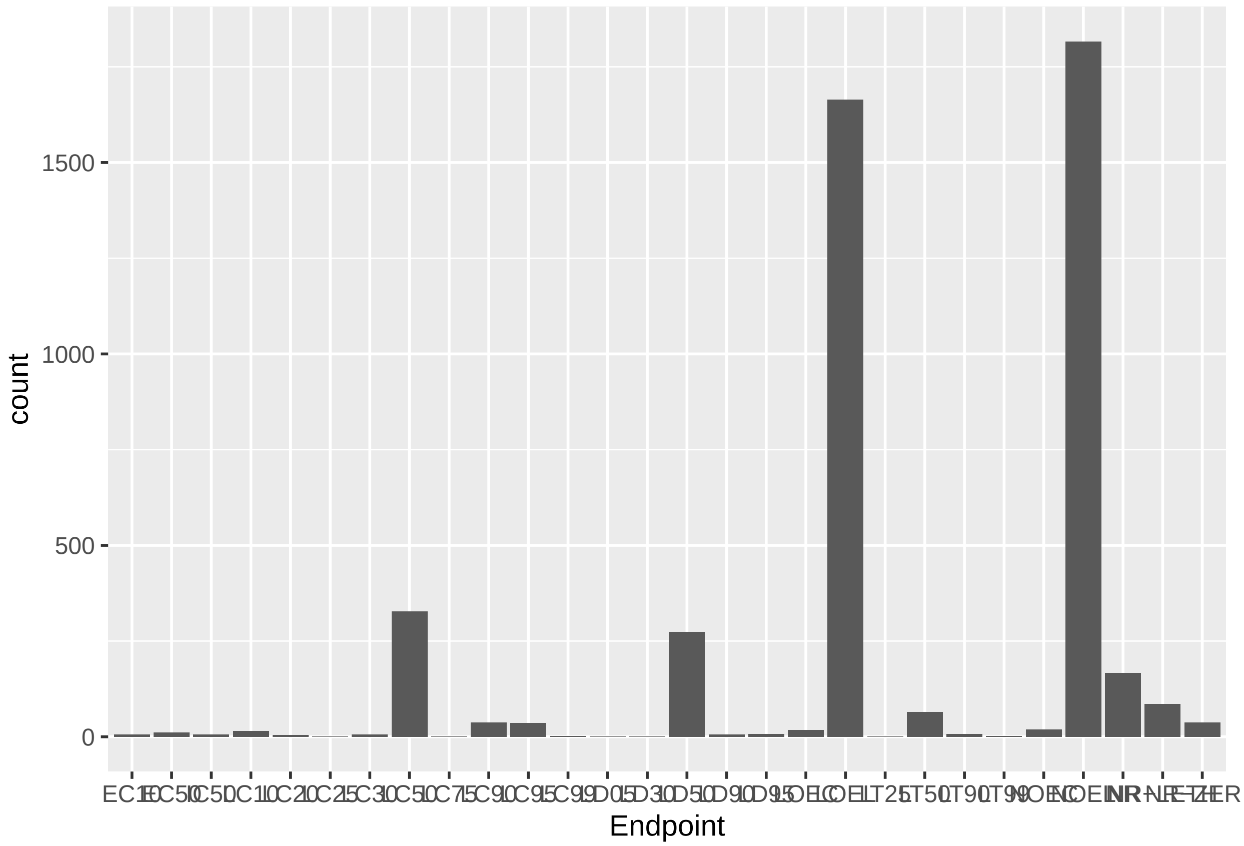


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are 'labs', followed by 'field natural', followed by 'field artificial'. The general trend shows that tests increase over time, with lab test locations peaking around 2015. The field natural test location on the other hand peaks around 2009, and then drops off after that.

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()
```



Answer: The two most common endpoints are 'NOEL' and 'LOEL'. 'NOEL' stands for 'no observable effect level', which means the highest dose given without producing a significant effect compared to the control. 'LOEL' stands for 'Lowest observable effect level', which means lowest dose producing effects that were significantly different from control.

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"
```

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

class(Litter$collectDate)
```

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

```
View(Litter)

unique(Litter$collectDate)
```

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

Answer: Using the 'unique' function, I found out that litter was sampled on 08/02/2018, and 08/30/2018

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`?

```r
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
```

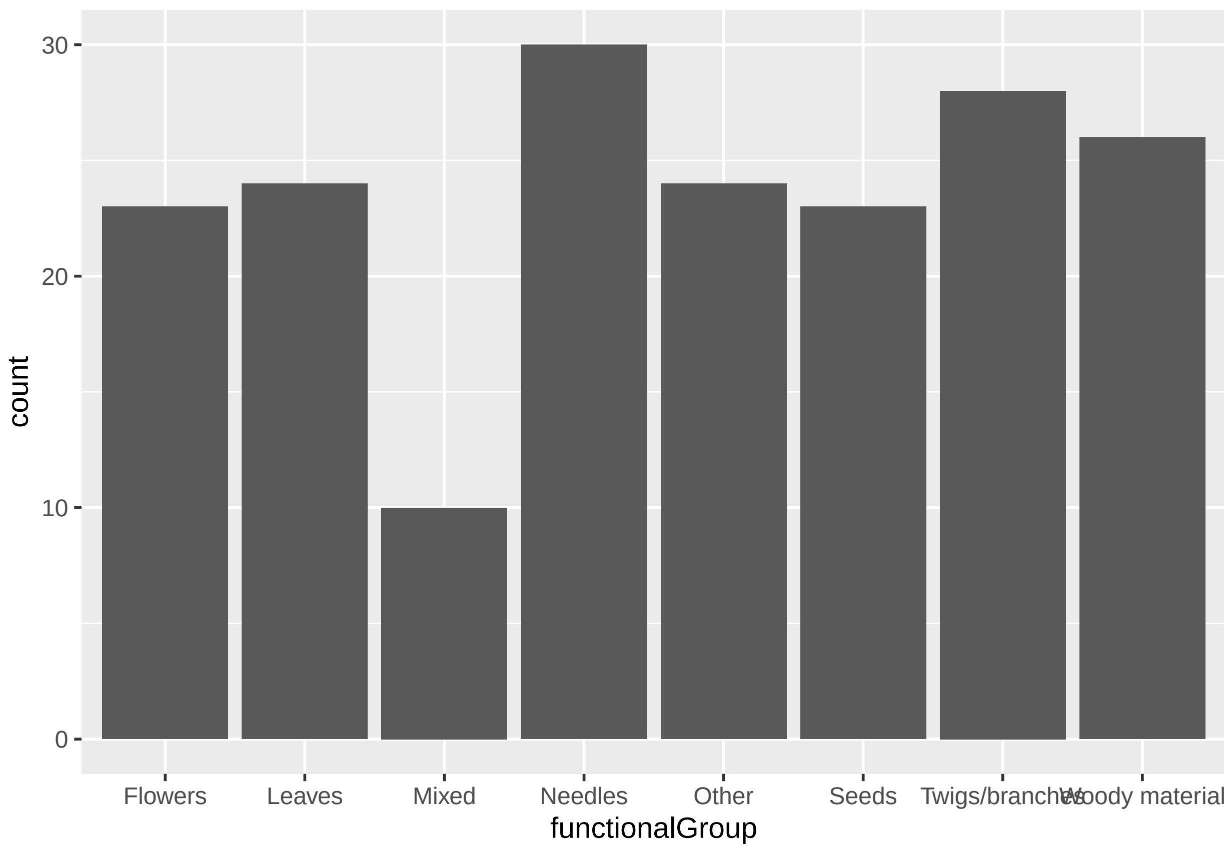
```r
summary(Litter$plotID)
```

```
## 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 plots. Unique is different from summary because unique only gives you unique plots at Niwot Ridge. Summary gives you unique plots but also gives you the number of samples from each plot.

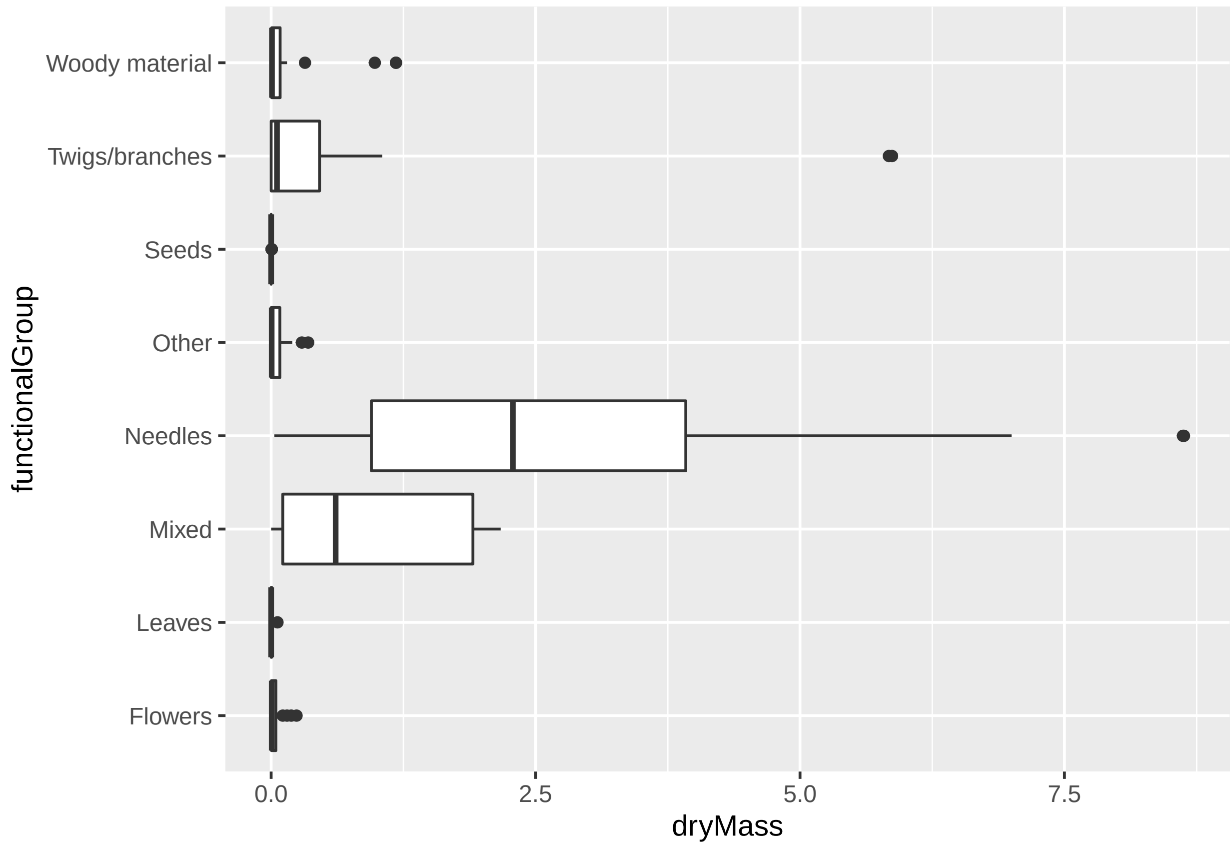
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

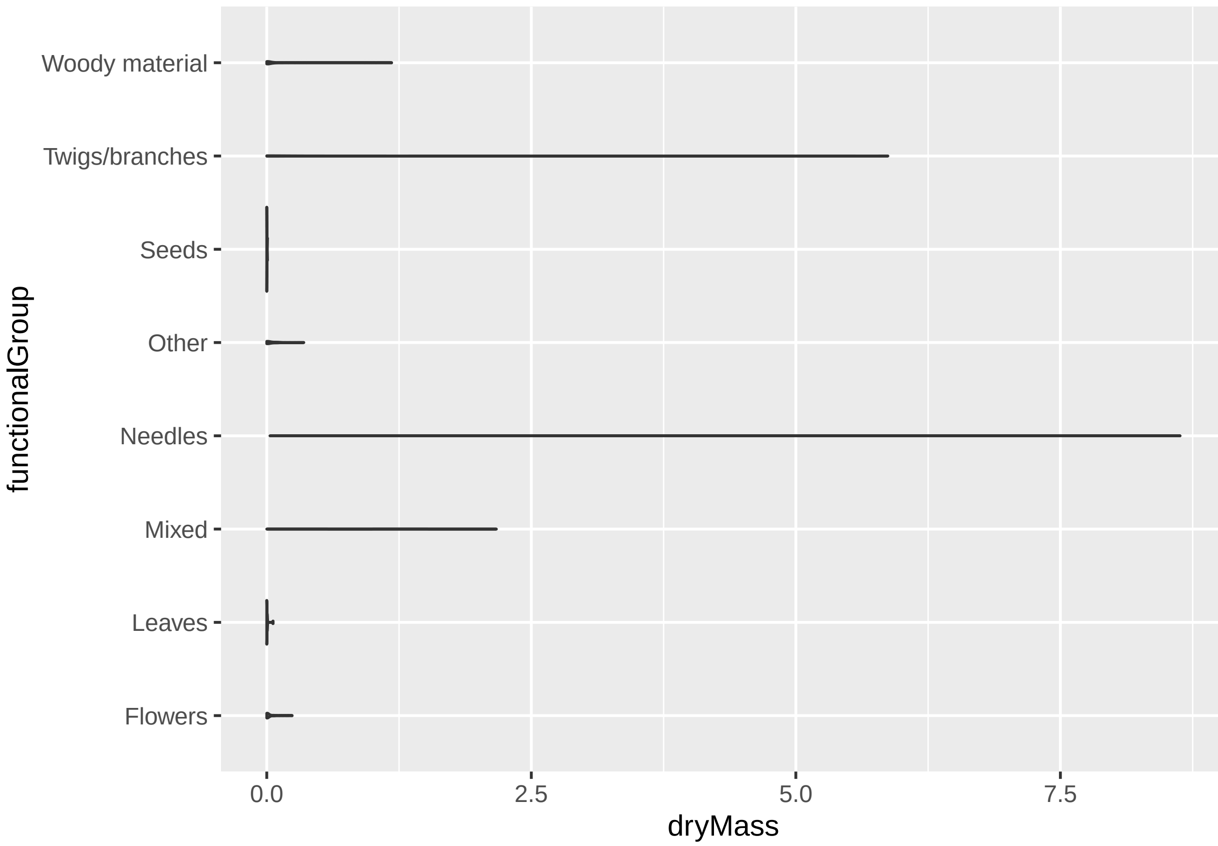


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

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




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



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

> Answer: Boxplot is more effective because in this case, the median for the data is very close to zero for most functional groups, and is not represented well in violin plots.

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

> Answer: Needles has the highest biomass, followed by mixed.