Assignment 3: Data Exploration

Benjamin Culberson, 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.**

```
library(tidyverse)
#getwd()
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: There have been several studies on neonicotinoids that linked the use of these chemicals to honey-bee collapse disorder. As environmentalists and data scientists, we might be interested in doing some of research of our own on this topic. If neonicotinoids are really that harmful to bees, the entire environmental field should be concerned.

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: Woody debris and litter can mde used as a proxy for the health of a forest. This debris and litter allows for the recyling of nutrients in the forest and therefore improve biodiversity. As a result, the measure of litter and woody debris can be used as information regarding the biodiversity of a 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: *Litter is defined as material that is dropped from the forest canopy and has a butt end diameter <2cm and a length <50 cm, this material is collected in elevated 0.5m^2 PVC traps* Fine wood debris is defined as material that is dropped from the forest canopy and has a but end diameter <2cm and a length >50 cm, this material is collected in ground traps as longer material is not reliably collected by the elevated traps *Litter and fine woody debris sampling is executed at terrestrial NEON sites that contain woody vegetation >2m tall

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
dim(Neonics)
```

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

   Answer: There are 4623 observations of 30 variables

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?

```r
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 and mortality by a long shot. These effects might specifically be of interest becaues we're interestd in the effect the neonicotinoids have on the health of bee colonies. These colonies aren't acting strangely, they are dying at a concerning rate and we need to know if these neonicotinoids are the cause.

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.

```r
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: It appears that the Honey Bee, the Parasitic Wasp, the Buff Tailed Bumblebee, the

Carniolan Honey Bee, the Bumble Bee, and the Italian Honeybee are the most commonly studied species in the dataset. All these species pollinate most of our agricultural crops and therefore are of most economic value to humanity. The loss of the mosquito might be exciting. The loss of all of the world's bees could mean disaster.

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.)
```

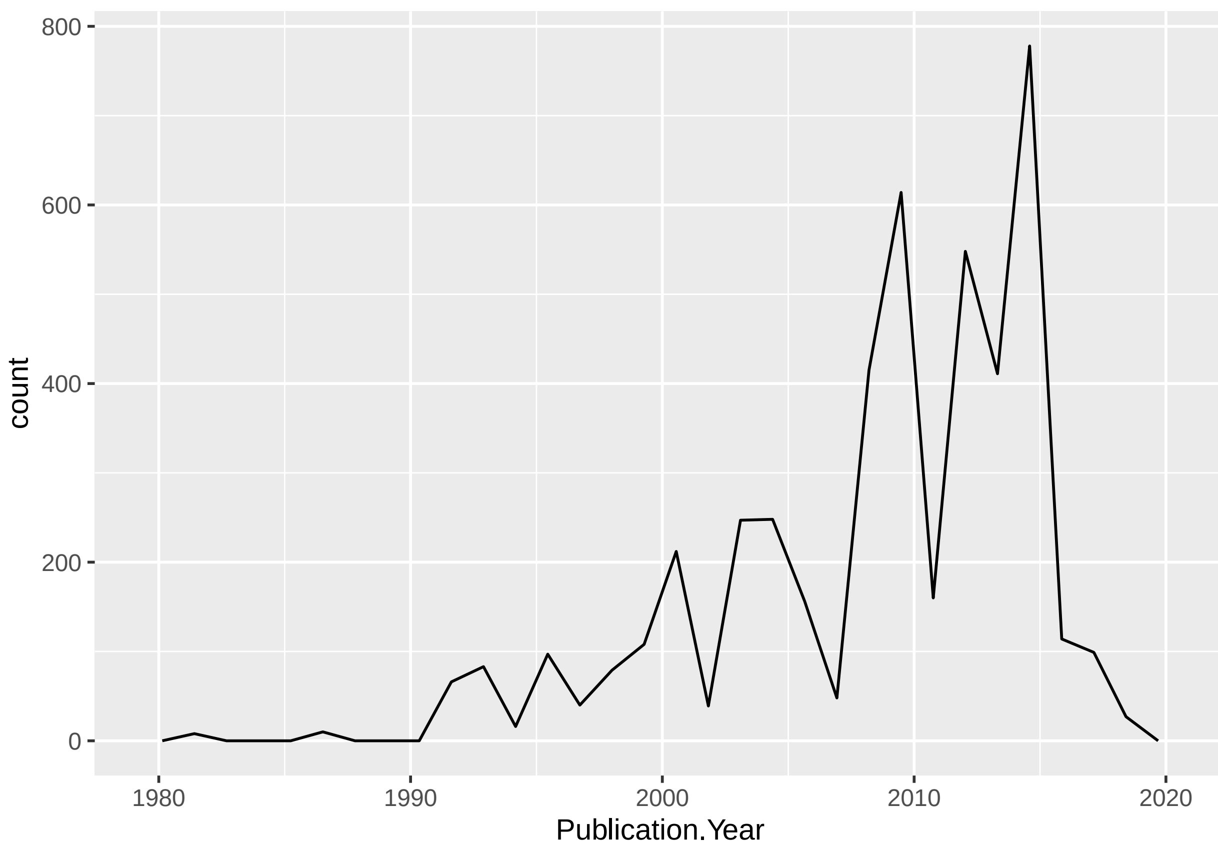
```
## [1] "factor"
```

Answer: The class of Conc.1..Author is "factor" and not numeric becaue when we read the data into R, there were some non-numeric values in the column. As a result, R saw that the column was not entirely numeric so it treated it as a factor.

Explore your data graphically (Neonics)

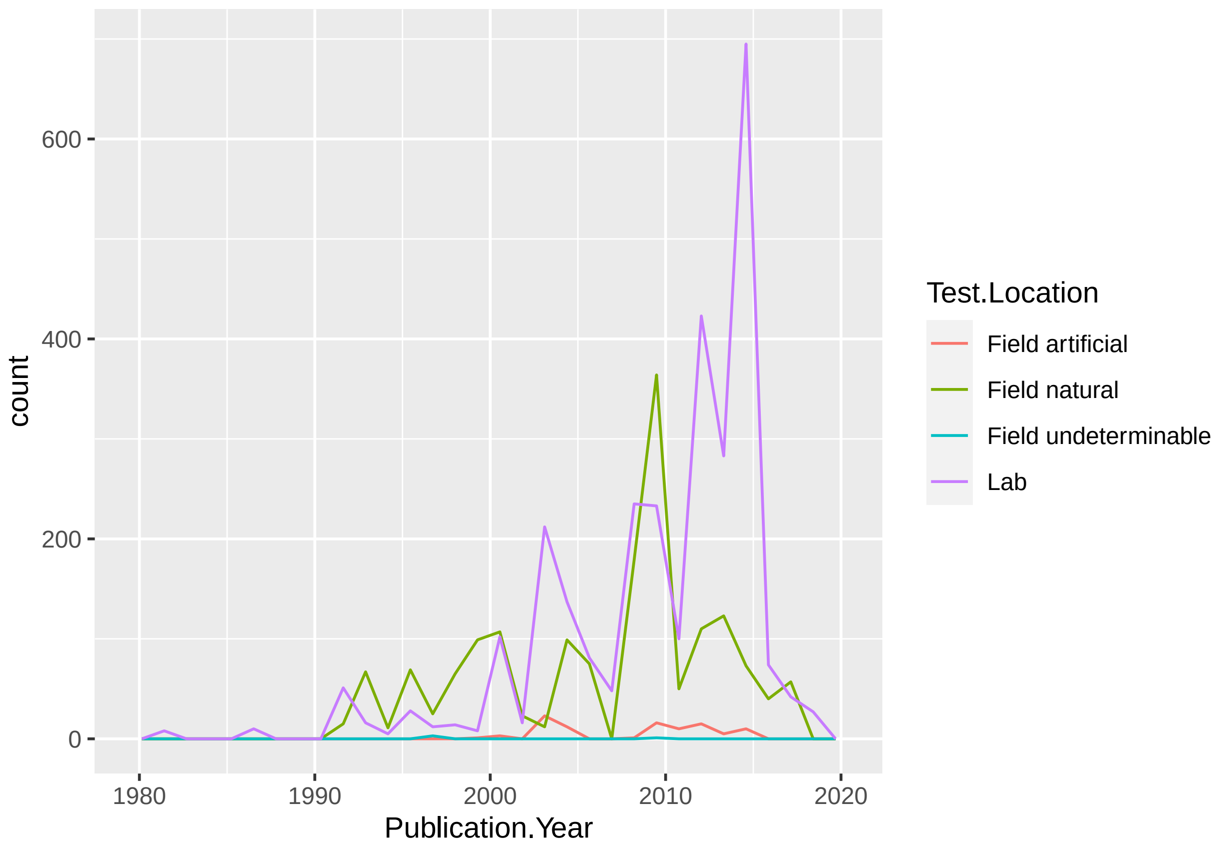
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 30)
```



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

```r
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins = 30)
```

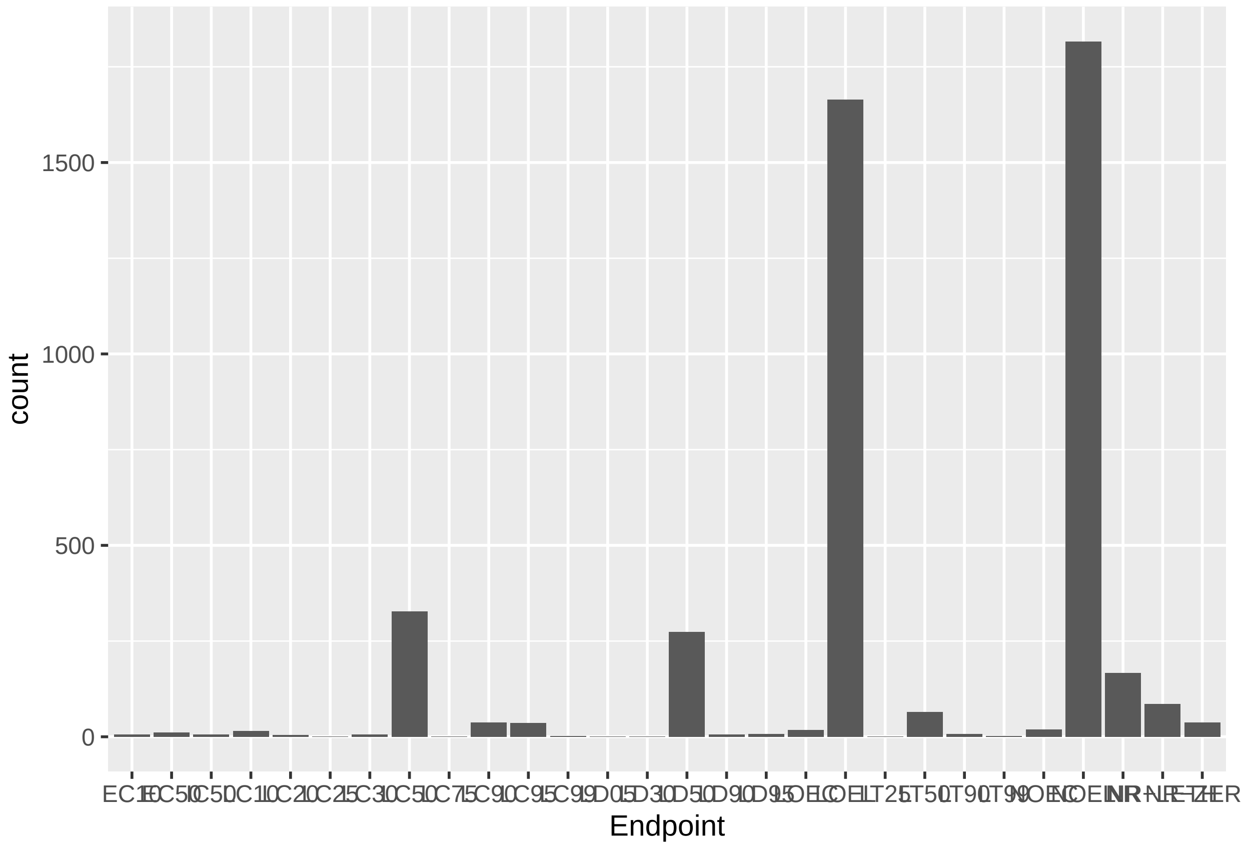


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: It appears that the most common test locations were "Field natural" and "Lab". They also do appear to differ over time. "Field natural" was the most common test location in 2008-2010, but a laboratory setting was always rather popular and become very popular from 2011-2015.

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.

```r
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar()
```




```r
summary(Neonics$Endpoint)
```

```
##    EC10    EC50    IC50    LC10    LC20    LC25    LC30    LC50    LC75    LC90
##       6      11       6      15       5       1       6     327       1      37
##    LC95    LC99    LD05    LD30    LD50    LD90    LD95    LOEC    LOEL    LT25
##      36       2       1       1     274       6       7      17    1664       1
##    LT50    LT90    LT99    NOEC    NOEL      NR NR-LETH NR-ZERO
##      65       7       2      19    1816     167      86      37
```

> Answer: The two most common end points are NOEL and LOEL. They are defined as NOEL: no-observable-effect-level: Highest dose (concentration) producing effects not significantly different fomr response of constrol according to author's reported statistical test; and LOEL: Lowest-observed-effect-level: lowest dose (concentration) producing effects that were significantly different (as resported by authors) from resoponses 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.

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

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

```r
unique(Litter$collectDate)
```

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

Answer: Litter was sampled on the 2nd and 30th day of August 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$siteID)
```

```
## [1] NIWO
## Levels: NIWO
```

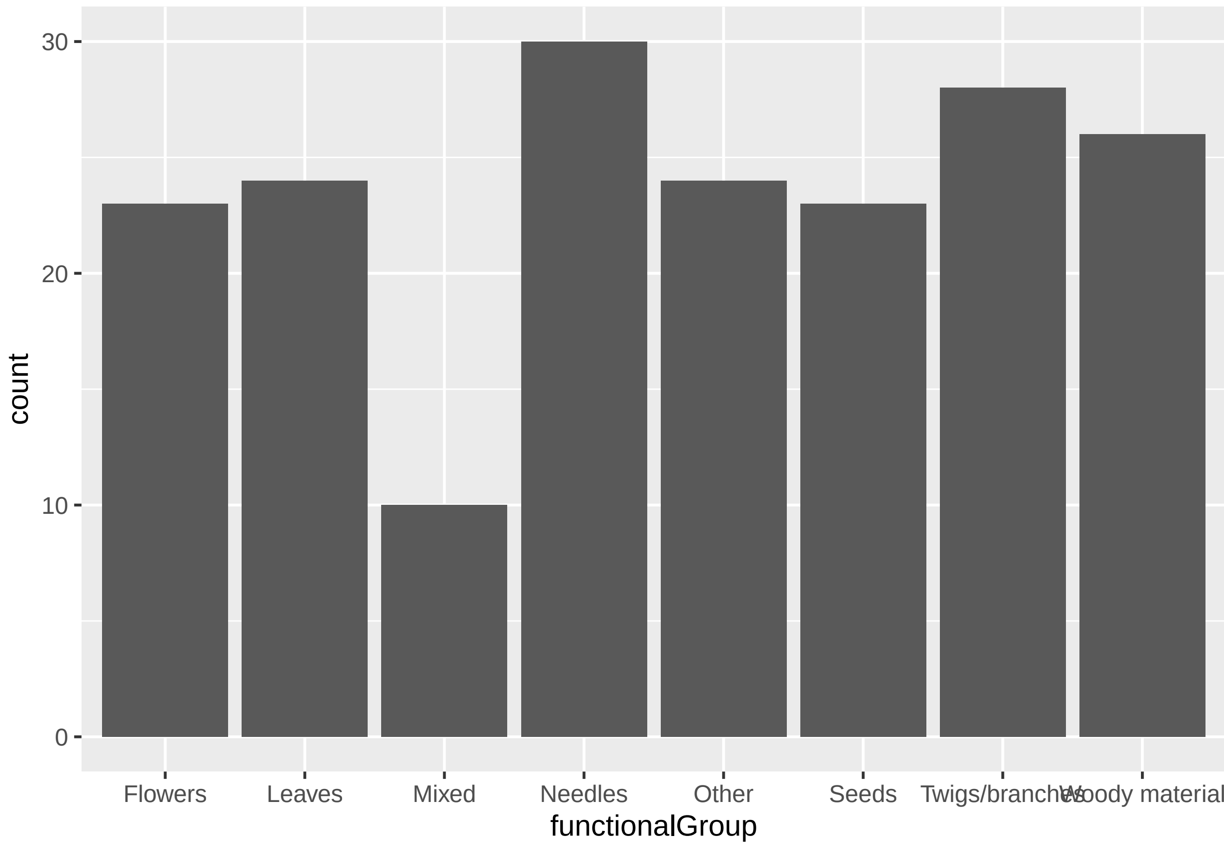
```r
summary(Litter$siteID)
```

```
## NIWO
##  188
```

Answer: There were 188 plots sampled at Niwot Ridge. The information obtained from 'unique' is different from that obtained from 'summary' because 'unique' simply removes any duplicates from the data frame it's iteratinng over. 'summary' provides a count of the data points at each site. 'unique' is sufficient to determine how many plots were sampled at Niwot ridge however, because if 'unique' only returns one value and there are 188 observations in this data frame, then there must be 188 plots at Niwot ridge.

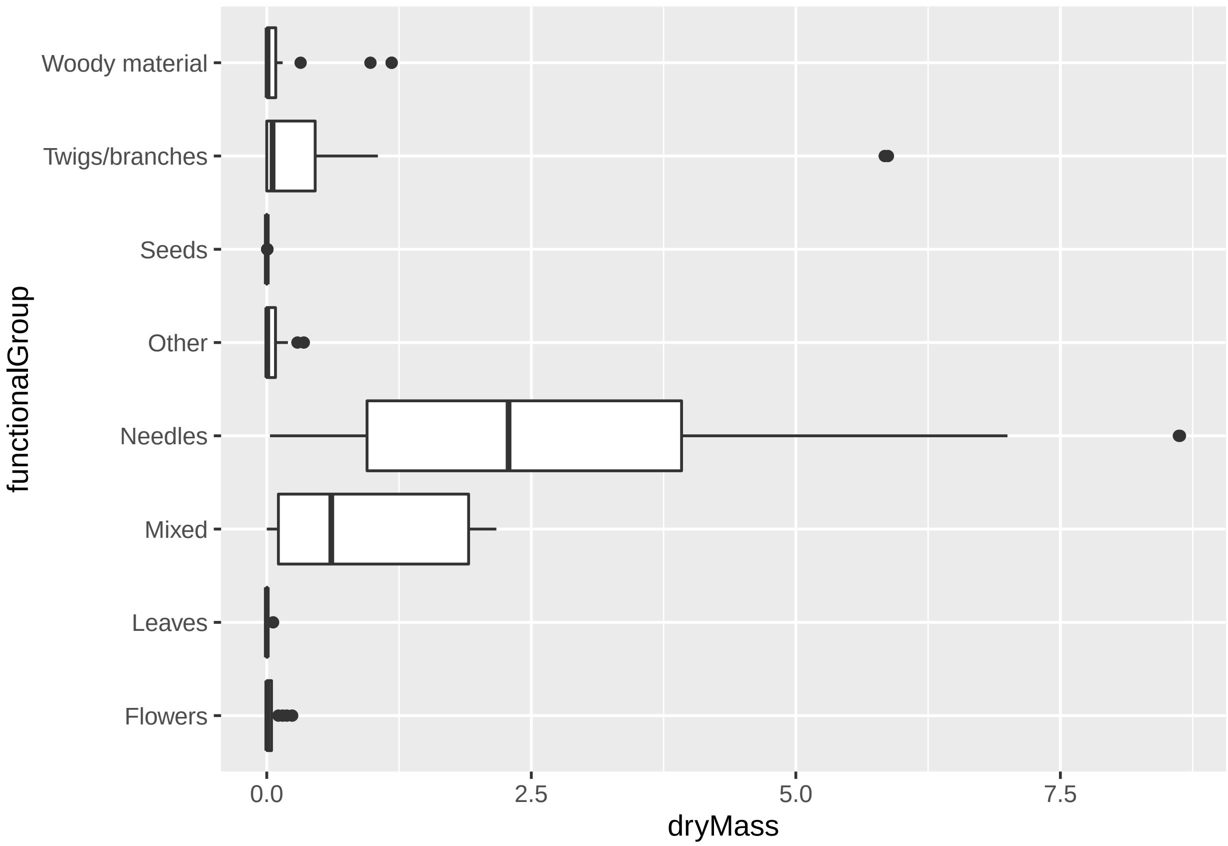
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```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.

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

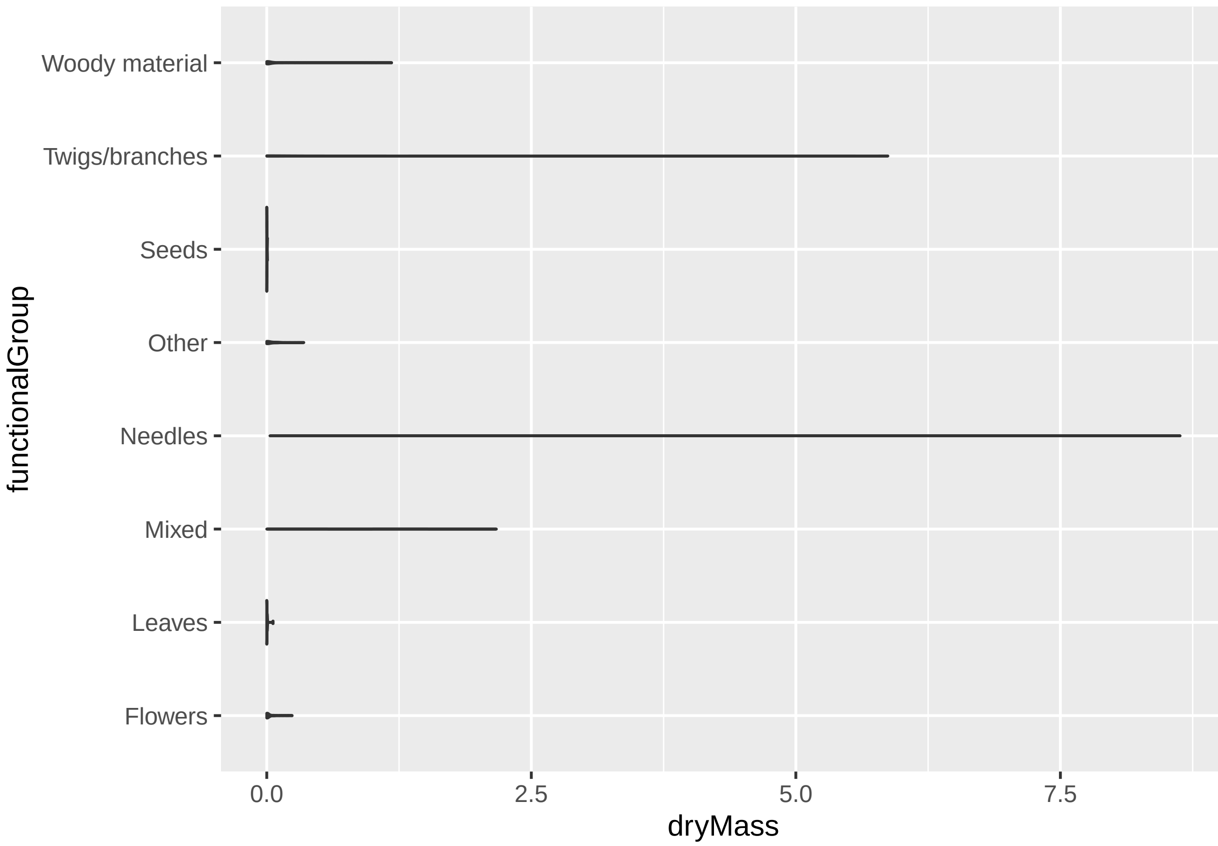



```r
ggplot(Litter) +
  geom_violin(aes(x = dryMass, y = functionalGroup),
              draw_quantiles = c(0.5))
```

```
## 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: A box plot is more effective at vizualization than a violin plot in this case because there are too few dryMass values. There is no need or possibility to demonstrate the conenctration of data point at various dryMass values because there simply aren't enough data points at that value within each functionalGroup for the violin plot to function.

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

> Answer: At these sites, it seems that "Needles" have the most biomass at these sites. However it should be said that "Mixed" and "Twigs/Branches" have more mass than the remaining other types of litter.