Assignment 3: Data Exploration

Changxin Yu

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A03_DataExploration.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change "Student Name" on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.
6. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai.

The completed exercise is due on Sept 30th.

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 include the subcommand to read strings in as factors.

```
getwd()
```

```
## [1] "E:/things/Duke University/study/2022 Fall/ENVIRON 872/EDA-Fall2022/Assignments"
```

```
setwd("E:/things/Duke University/study/2022 Fall/ENVIRON 872/EDA-Fall2022")
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 ecotoxicology of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

Answer: Some researches showed that neonicotinoids have negative effects on bees, insect-eating birds, aquatic invertebrates, etc. Therefore, it is necessary to study the ecotoxicology of neonicotinoids on insects for helping decide whether neonicotinoids can be used as insecticides or not.

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: Forest litter and woody debris are important part of forest ecosystem for their roles in carbon cycling, nutrient cycling and soil stabilization. To characterize this ecosystem component, studying litter and woody debris is essential.

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: 1. Sampling is at both temporal and spatial scales. 2. For spatial sampling, locations of tower plots are selected randomly near the airsheds. The trap layout in the plots could be targeted or randomized according to vegetation. 3. For temporal sampling, ground traps are sampled once a year, while the frequency of target traps sampling varies with the vegetation at the site.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

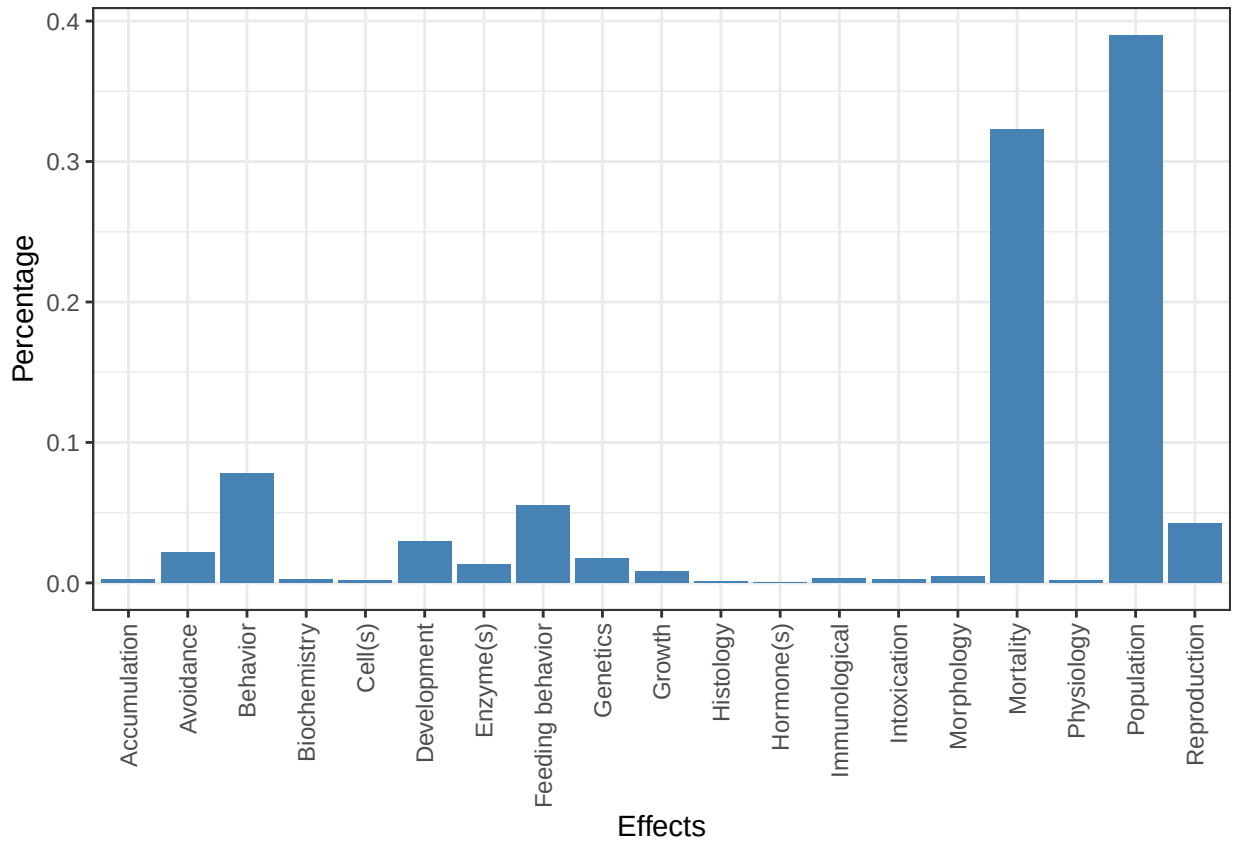
6. Using the `summary` function on the "Effect" column, determine the most common effects that are studied. Why might these effects specifically be of interest?

```
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
```

```
perc_eff <- as.data.frame(prop.table(summary(Neonics$Effect)))
perc_eff <- cbind(rownames(perc_eff), data.frame(perc_eff, row.names=NULL))
colnames(perc_eff) <- c("Effects", "Percentage")

ggplot(perc_eff, aes(x=Effects, y=Percentage))+theme_bw()+
  geom_bar(stat = 'identity', fill = 'steelblue')+
  theme(axis.text.x = element_text(angle=90, vjust=.5,hjust=1))
```



Answer: The most common studied effects are population and mortality. The study of neonicotinoids focus on the ecotoxicology on insects, so population and mortality are the two main parts of this study.

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.

```
rev(sort(summary(Neonics$Species.Common.Name)))[1:7]
```

```
##               (Other)             Honey Bee        Parasitic Wasp
##                   670                   667                   285
## Buff Tailed Bumblebee   Carniolan Honey Bee           Bumble Bee
##                   183                   152                   140
##      Italian Honeybee
##                   113
```

Answer: The six most commonly studied species are honey bee, parasitic wasp, buff tailed bumblebee, carniolan honey bee, bumble bee and italian honeybee, which are all species of bees. The reason that bees are of interest is that neonicotinoids have a worse effect on bees than other insects. Neonicotinoids are used widely as insecticides and can be absorbed by plants and then be kept in pollen and nectar, where bees come for food. And researches showed that some of neonicotinoids are toxic to bees when up to a certain dose.

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.Type..Author.)
```

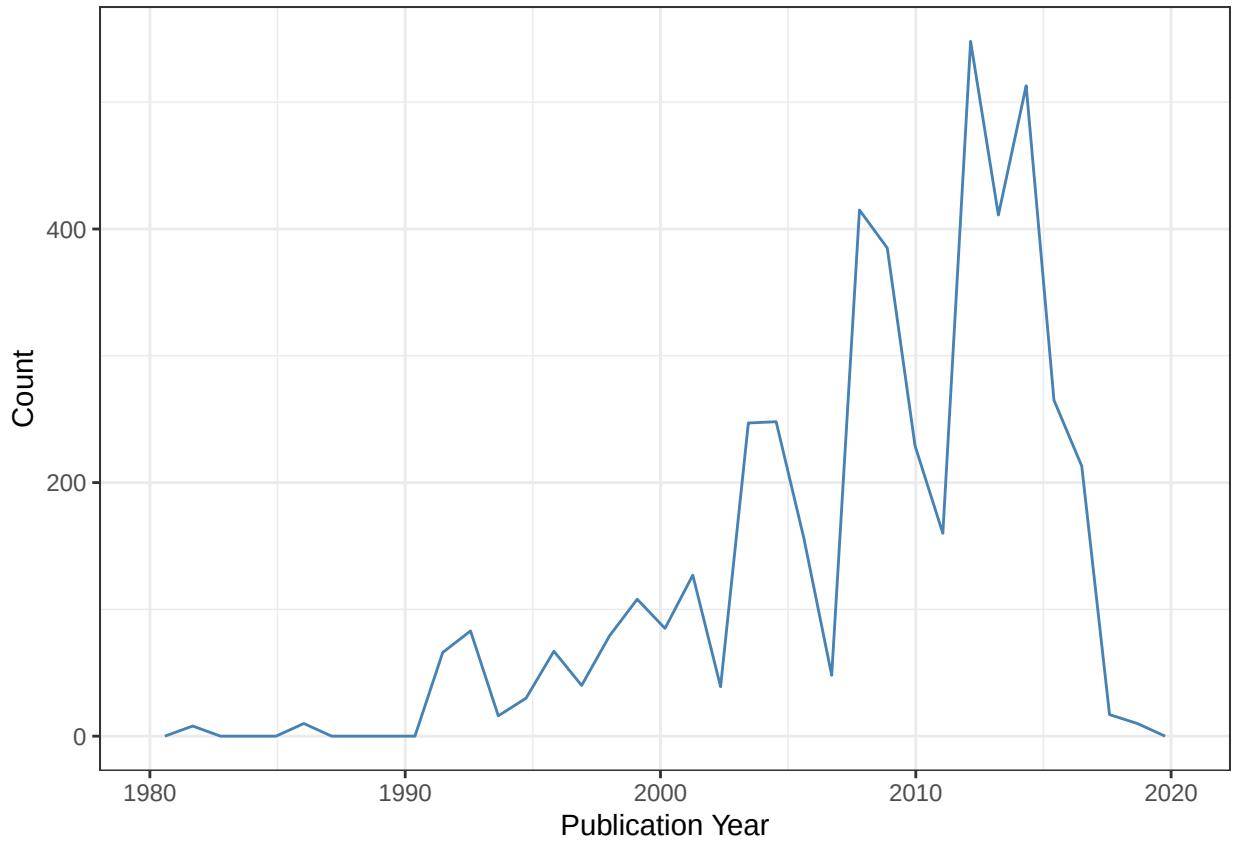
```
## [1] "factor"
```

Answer: It's not numeric because there are also some letters and symbols in this column. When importing data, it was identified as characters and was converted to factors.

Explore your data graphically (Neonics)

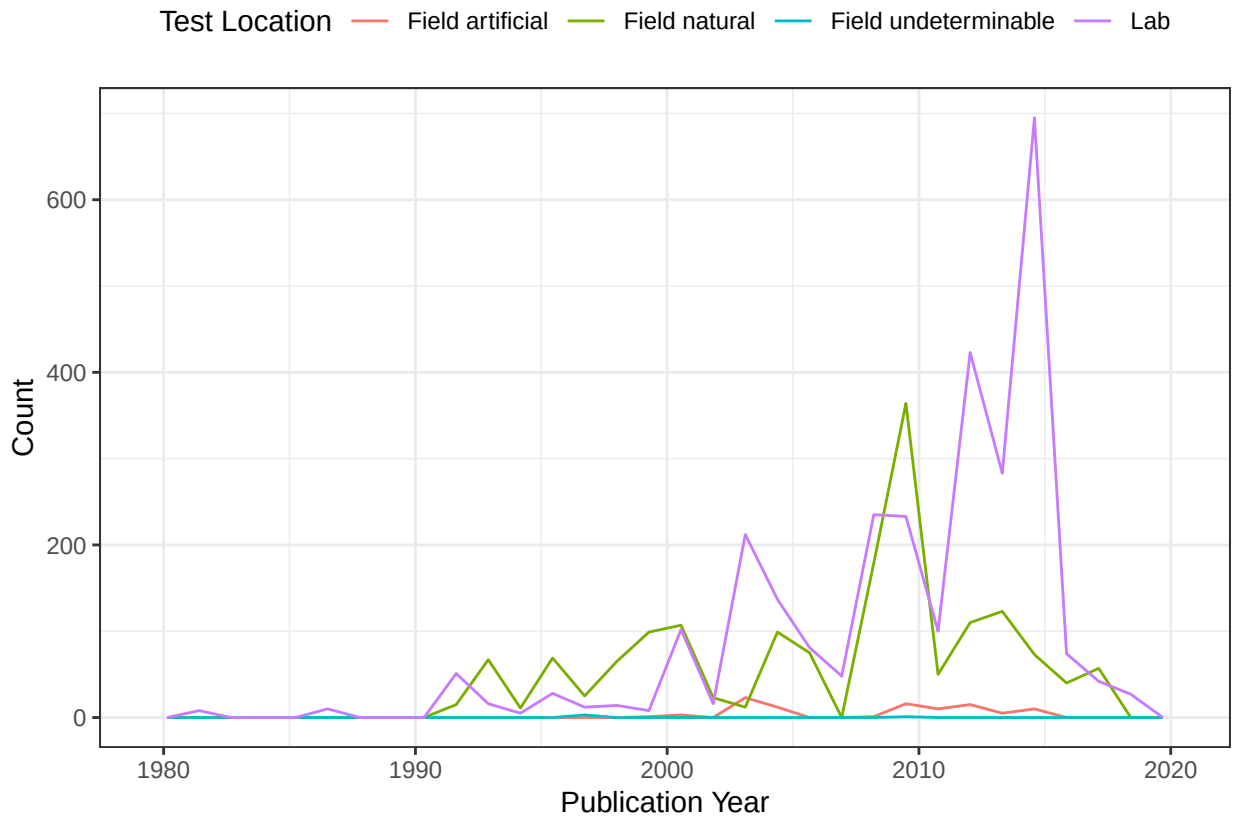
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics, aes(Publication.Year))+theme_bw()+
  geom_freqpoly(color="steelblue", bins=35)+
  labs(x='Publication Year',y='Count')
```



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

```
ggplot(Neonics, aes(Publication.Year, color=Test.Location))+theme_bw()+
  geom_freqpoly(bins=30)+
  labs(x='Publication Year',y='Count')+
  scale_color_discrete(name="Test Location")+
  theme(legend.position = "top")
```

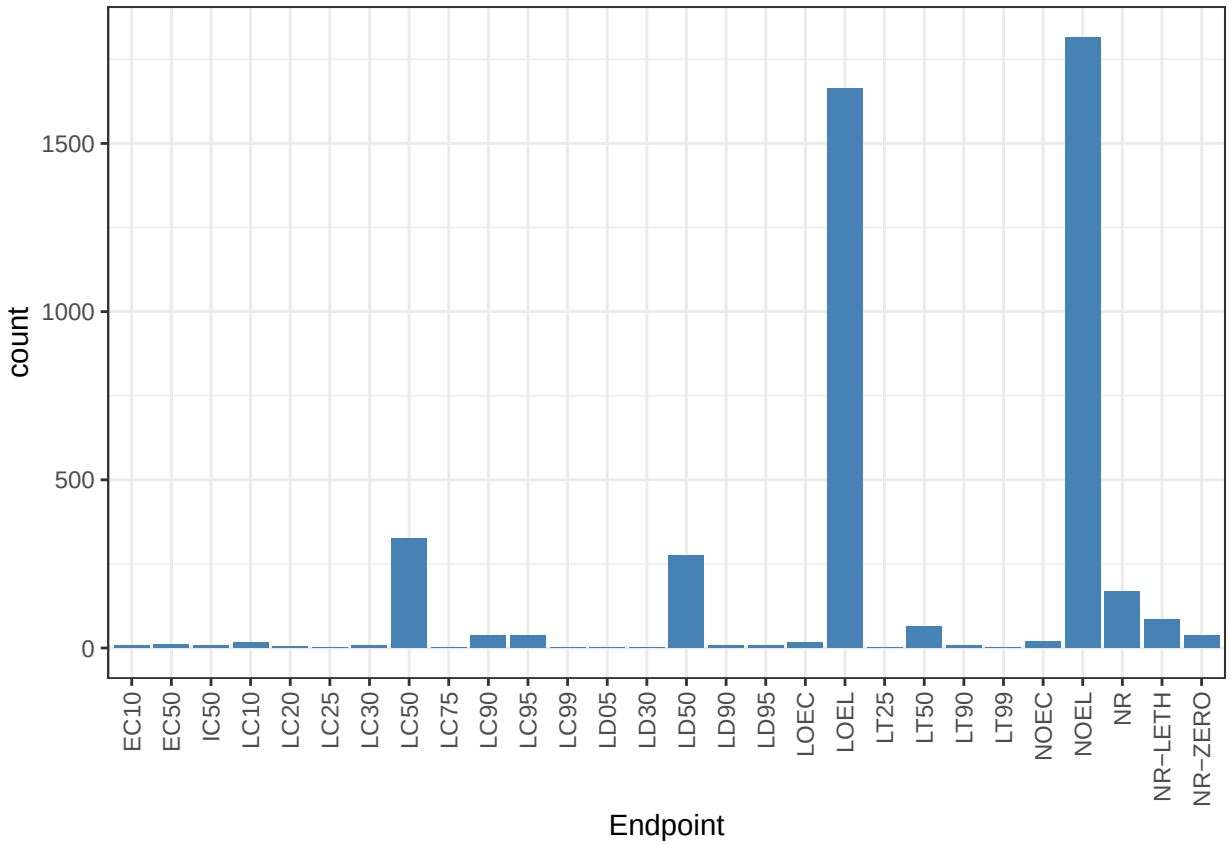


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test location is lab. Test locations do differ over time. There were lots of field natural research around 2009, while around 2014, there was much more laboratory research than before and after.

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(Endpoint))+theme_bw()+
  geom_bar(stat = 'count', fill = 'steelblue')+
  theme(axis.text.x = element_text(angle=90, vjust=.5,hjust=1))
```



Answer: The two most common end points are NOEL and LOEL. The NOEL is defined as no-observable-effect-level, which means that highest dose producing effects were not significantly different from responses of controls. While the LOEL is defined as lowest-observable-effect-level, which means that lowest dose producing effects 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.

```
class(Litter$collectDate)
```

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

```
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"
```

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
```

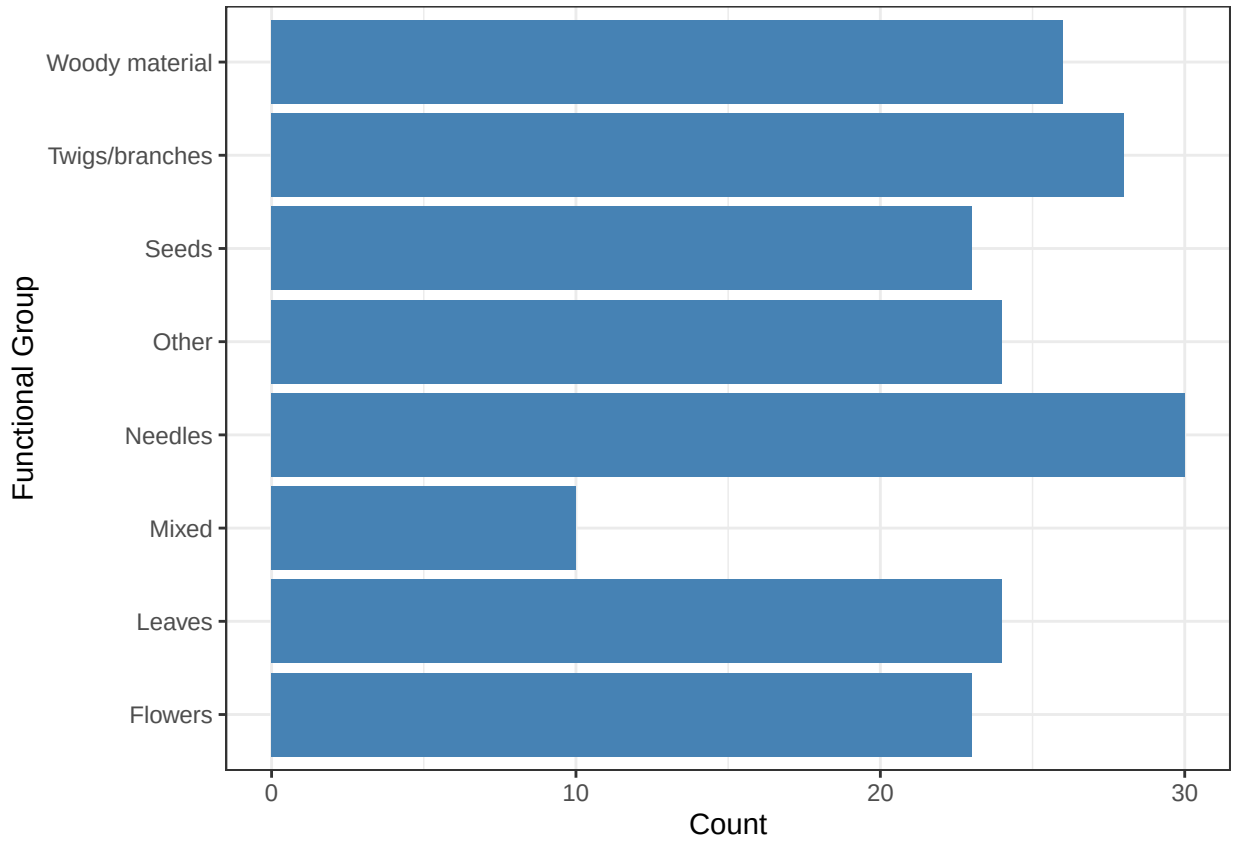
```
length(unique(Litter$plotID))
```

```
## [1] 12
```

Answer: There are 12 plots at Niwot Ridge. Function unique shows all unique elements. While function summary shows not only unique elements but also the counting of them.

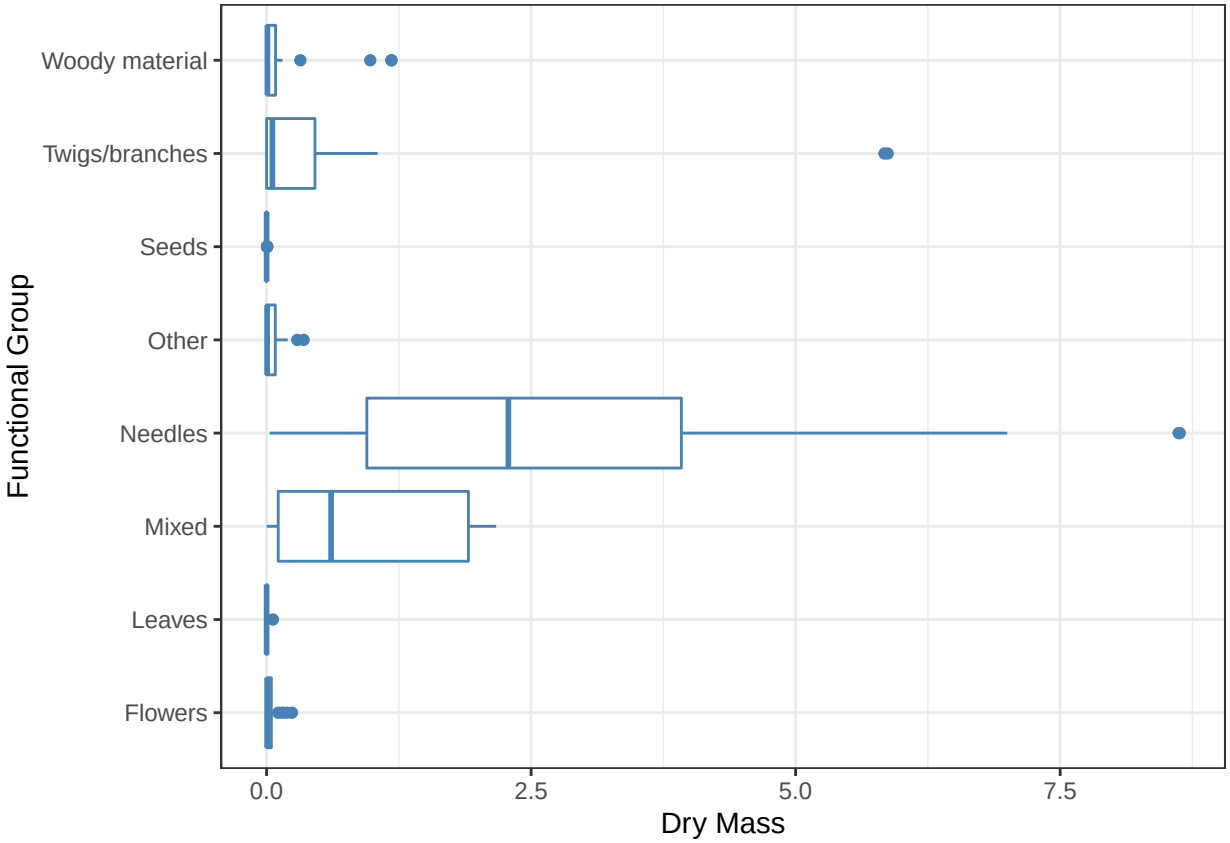
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(functionalGroup))+theme_bw()+
  geom_bar(stat = 'count', fill = 'steelblue')+
  labs(x='Functional Group',y='Count')+coord_flip()
```

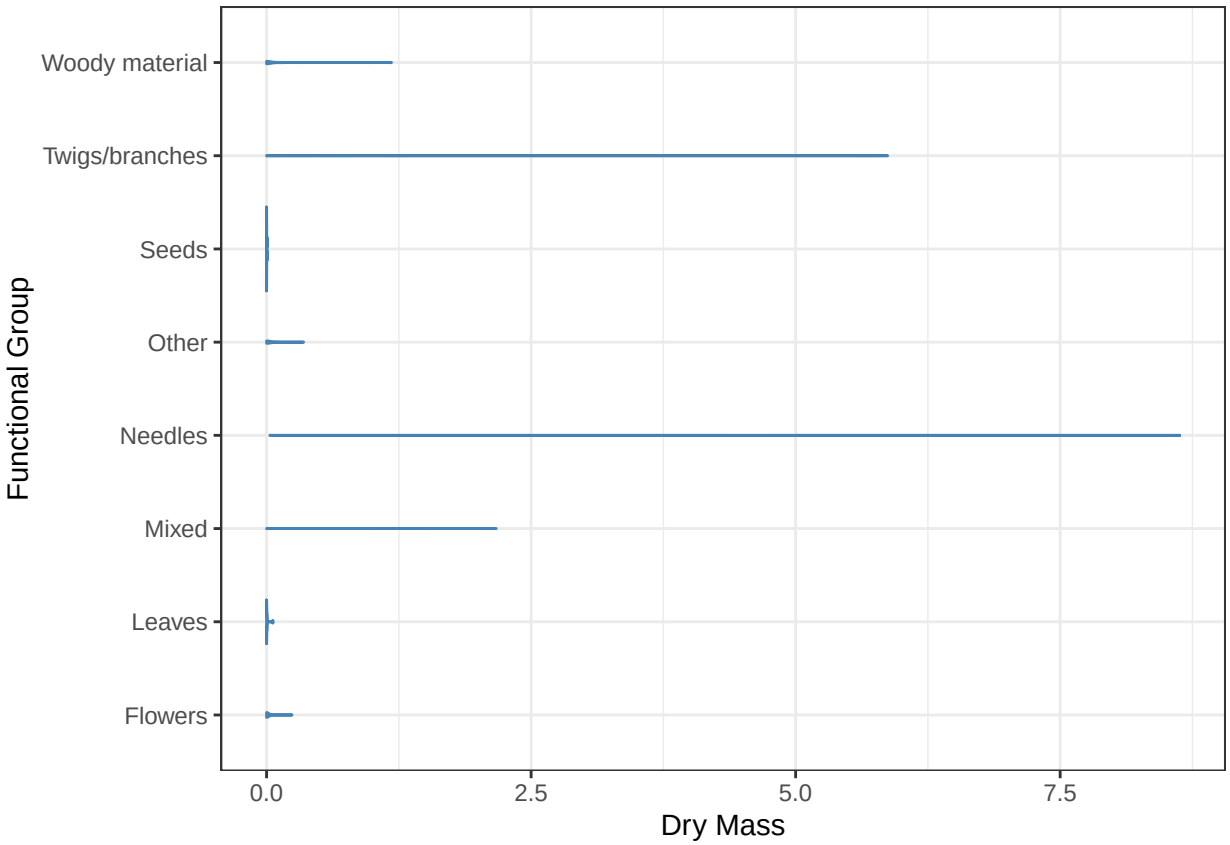


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(Litter,aes(x=functionalGroup,y=dryMass))+
  geom_boxplot(color="steelblue")+theme_bw()+
  labs(x='Functional Group',y='Dry Mass')+
  coord_flip()
```




```
ggplot(Litter,aes(x=functionalGroup,y=dryMass))+
  geom_violin(color="steelblue")+theme_bw()+
  labs(x='Functional Group',y='Dry Mass')+
  coord_flip()
```



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

Answer: In this case, boxplot is more informative with quartiles and outliers. Violin plot shows that the distribution of dry mass is fairly equal in each functional group.

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

Answer: Needles have the highest biomass according to the median.