Assignment 3: Data Exploration

Ana Andino

Fall 2024

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. Assign a useful **name to each code chunk** and include ample **comments** with your code.
5. Be sure to **answer the questions** in this assignment document.
6. When you have completed the assignment, **Knit** the text and code into a single PDF file.
7. After Knitting, submit the completed exercise (PDF file) to the dropbox in Canvas.

TIP: If your code extends past the page when knit, tidy your code by manually inserting line breaks.

TIP: If your code fails to knit, check that no `install.packages()` or `View()` commands exist in your code.

Set up your R session

1. Load necessary packages (tidyverse, lubridate, here), check your current working directory 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.

```
#uploading r packages
#install.packages('tidyverse')
#install.packages('lubridate')
#install.packages('here')

#Check working directory
getwd()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```
library(here)
```

```
## here() starts at /home/guest/EDE_Fall2024
```

```
here()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```
# Relative file path. This is recommended instead of the absolute path
Neonics <- read.csv(
  file = here("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"),
  stringsAsFactors = TRUE)

Litter <- read.csv(
  file = here("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv"),
  stringsAsFactors = TRUE)


#Import the files and now they are visible in the Environment
#read.csv(here("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"),           stringsAsFactors = TRUE)
#read.csv(here("./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: There are many reasons on why we could be particularly interested in neonicotinoids, especially when they can be a threat or a type of toxic that can now be found in water and soil samples in the US. Hence they have significant environmental and ecological impacts. For instnace: 1. There are studies that show that neonicotinoids have been linked to **declines in bee populations** which risks crop production and biodivertisty. 2. There is a concern that due to their extensive use, some insect populations have **developed resistance**. 3. **They can persist in the soil and water systems** which leads to expose to other insect populations. These are just a few examples on why we should study neonicotinoids.

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: There are many reasons on why we should be interested in studying litter and woody debris that falls to the ground in forest. Among them: 1. Forest litter - fallen leaves, branches - decompose over time which releases essential nutrients back to the soil. This becomes **nutrient cycling** as the process supports fertility and sustains plants. 2. Woody debris and litter

contribute to **storing capacity** of the forest as much gets stored in the soil. This becomes an important factor in mitigating climate change. 3. It protects the forect from soil erosion and reduces the impacts of heavy rain. 4. Debris and litter create microhabitats for many organism in the forest such as insects, fungi and some animals Thus, it supports **biodiversity**.

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. The sampling occurs only in tower plots (AD[06]) and the selection of them is random within the 90% flux footprint of the primary and secondary airsheds - as necessary to accommodate sufficient spacing between plots. 2. Litter is collected in elevated 0.5mˆ2 PVC traps which are 0.5mˆ2 square with mesh 'basket' elevated approximately 80cm above the ground. 3. Fine wood is collected in ground traps which are 3 m x 0.5 m rectangular areas.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#Getting dataset dimensions rows x columns
dim(Neonics)
```

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

```
dim(Litter)
```

```
## [1] 188  19
```

   Answer: The Neonics dataset has the following dimension: 4,623 x 30 The Litter dataset has the following dimension: 188 x 19

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? [Tip: The `sort()` command is useful for listing the values in order of magnitude...]

```
summary_effects <- summary(Neonics$Effect)
sort(summary_effects, decreasing = TRUE)
```

```
##       Population        Mortality         Behavior Feeding behavior
##             1803             1493              360              255
##     Reproduction      Development        Avoidance         Genetics
##              197              136              102               82
##        Enzyme(s)           Growth       Morphology    Immunological
##               62               38               22               16
##     Accumulation     Intoxication     Biochemistry          Cell(s)
##               12               12               11                9
##       Physiology        Histology       Hormone(s)
##                7                5                1
```

Answer: The top five most studied effects are population, mortality, behavior feeding behavior, reproduction, avoidance. Genetics, enzymes, growth, morphology, immunological, accumulation, intoxication, biochemistry, cells, physiology, histology and hormones con afters, respectively. I believe the answer to why there is an interest in studying these effects ties up with one of the questions above - why are we interested in the Neonics dataset. In general, it can be said that studying these effects might aid researchers and policymakers to make decisions about the right and correct use of neonicotinoids. This is more relevant when we go back to the negative spillover effects of neonicotinoids as it can help mitigate risks and promote sustainable practices that protect agricultural practices and biodiversity.

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.[TIP: Explore the help on the `summary()` function, in particular the `maxsum` argument...]

```
#I will use the maxsum argument  to focus on the sex top species studies.              Maxsum = 7 so t
summary(Neonics$Species.Common.Name, maxsum = 7)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                       285                   183
##   Carniolan Honey Bee            Bumble Bee      Italian Honeybee
##                   152                       140                   113
##                (Other)
##                  3083
```

Answer: Honey Bee (667), Parasitic Wasp Buff (285), Tailed Bumblebee (183),Carniolan Honey Bee (152), Bumble Bee (140), and Italian Honeybee (113)

8. Concentrations are always a numeric value. What is the class of `Conc.1..Author.` column in the dataset, and why is it not numeric? [Tip: Viewing the dataframe may be helpful...]

```
#Since I have two data sets, I need to specify from which data set comes the column      in question.

class(Neonics$Conc.1..Author.)
```

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

```
#View(Neonics)
```

Answer: It is enlisted as Factor. Even though the data has numbers, it is noticeable that it also bins numbers into categories (thus, a factor). For instance, we have numbers listers as $<$ or $>$. Also, there are numbers that have a "/" in the end which makes them not count numbers anymore.

Explore your data graphically (Neonics)

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

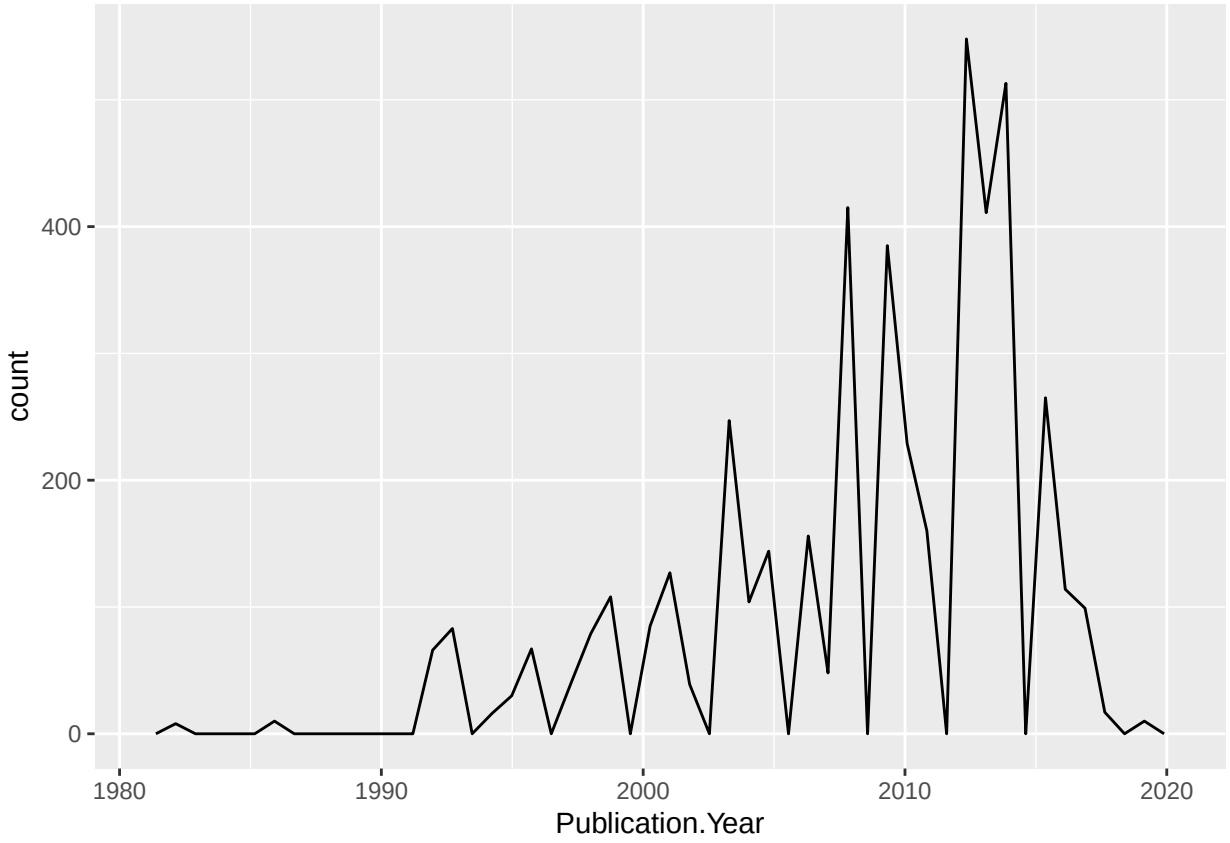
```r
class(Neonics$Publication.Year)
```

```
## [1] "integer"
```

```r
#Visualization for numerical value


#install.packages("ggplot2")
library(ggplot2)

#Create a graph using publication year in x axis with bins of 50
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50)
```



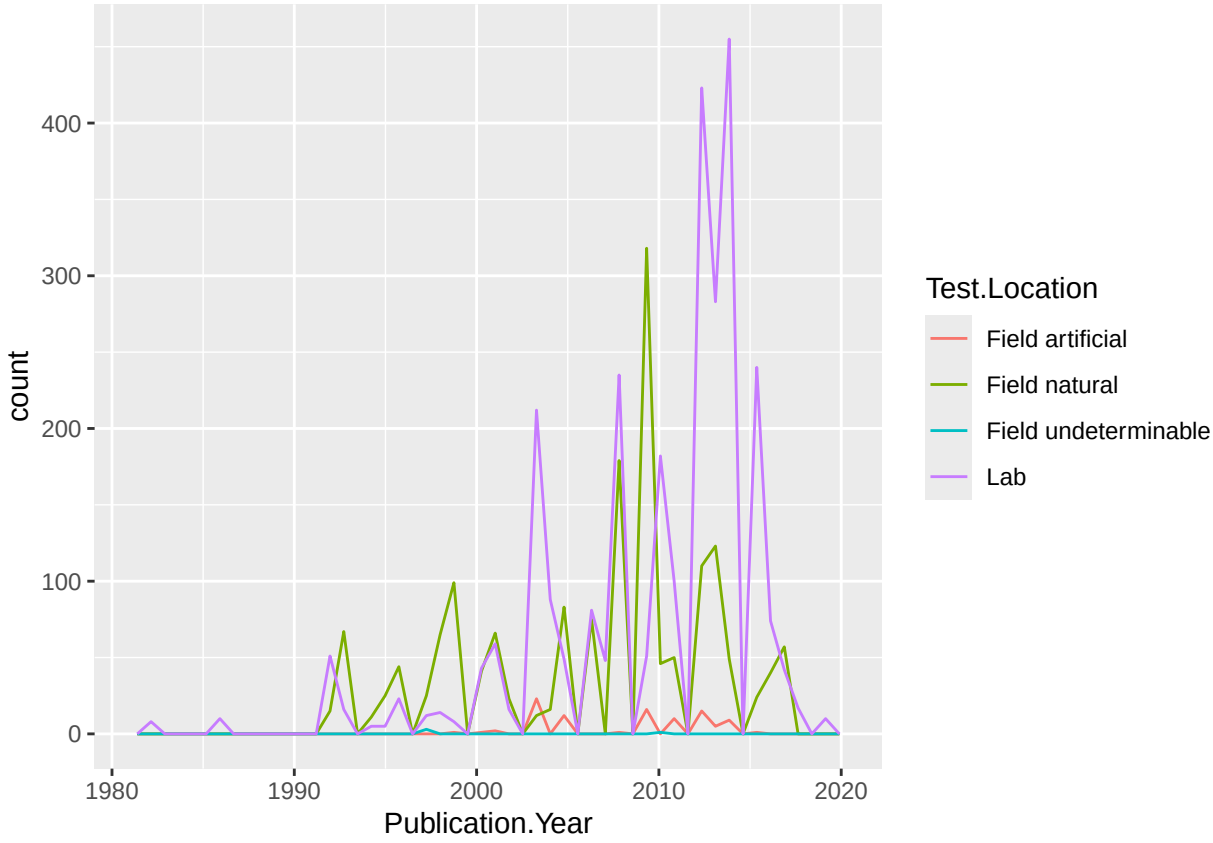
10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```r
#verify if the variable is a numeric variable
class(Neonics$Test.Location)
```

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

```
#same graph as before but now diffferentiating by test location

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



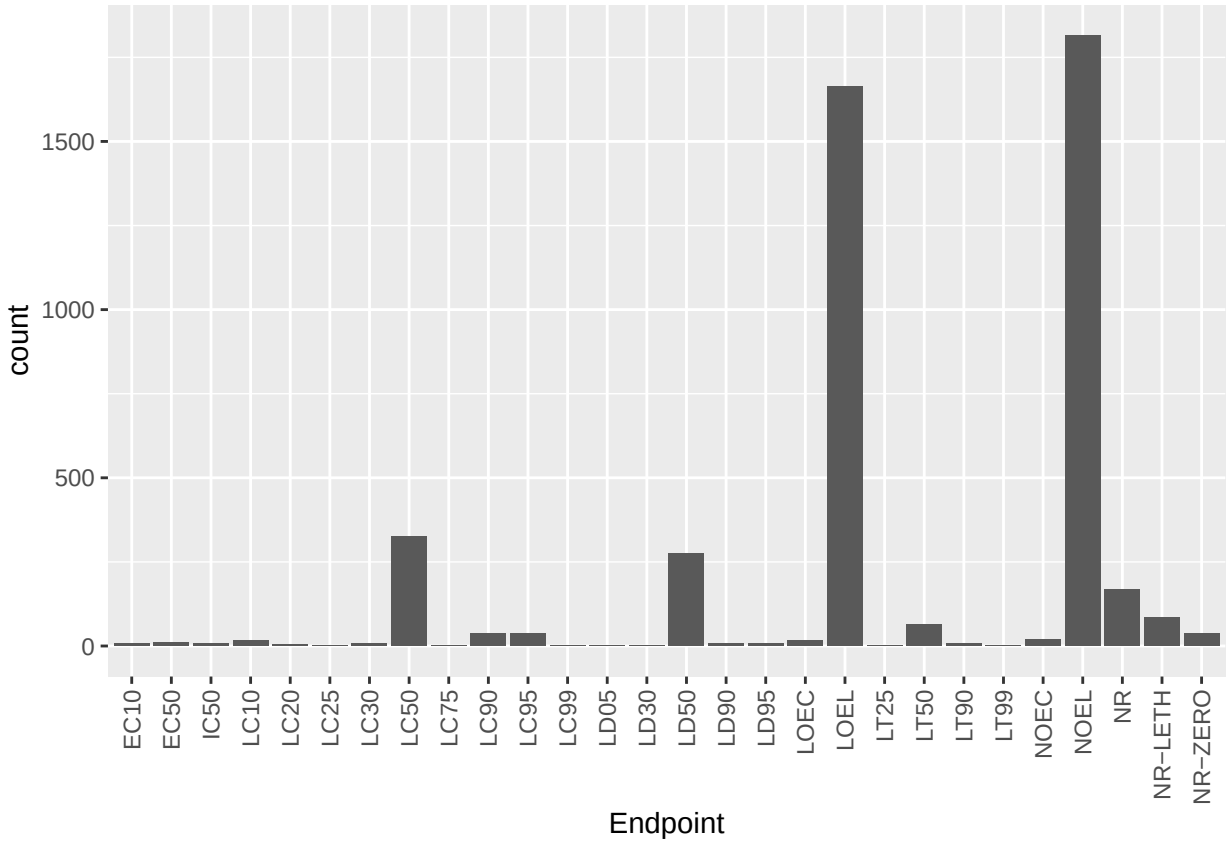
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The two most common test locations are labs and field experiment, with the former being more prominent. With time we can see that Lab locations are becoming more common even though there is a slight decrease in the last years. Also, we can see that between 2000 and 2010, field experiments were slightly more common than lab but this trend changed after 2012.

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.

[**TIP**: Add `theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))` to the end of your plot command to rotate and align the X-axis labels...]

```
ggplot(Neonics) +
  geom_bar(aes(x = Endpoint)) +
            theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer:The two most common endpoints are NOEL and LOEL. They are both Terrestial endpoints. LOEL is "the lowest observable effect level: **lowest dose** (concentration) producing effects that were significantly different from response of controls. While NOEL is"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.

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

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

```
#Class: Factor. Then, change to a date.

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

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

```
#Now collectDate has date format.
```

13. Using the `unique` function, determine how many different plots were sampled at Niwot Ridge. How is the information obtained from `unique` different from that obtained from `summary`?

```
unique(Litter$plotID) #Gives and array
```

```
##  [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
```

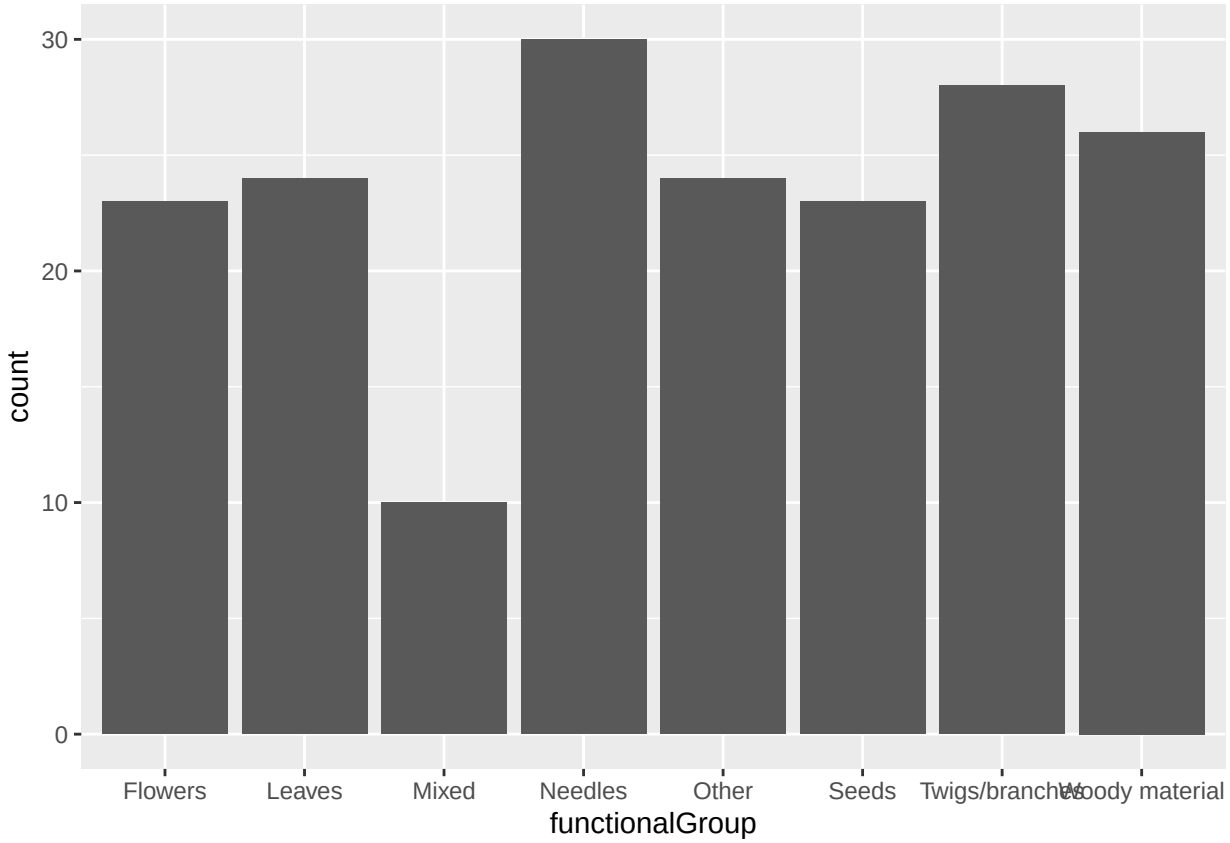
```
summary(Litter$plotID) #Gives the plots but also its counts.
```

```
## 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 different plots that were sampled at Niwot Ridge. The unique functions gives you the array of plots then specifying there are 12 levels, whereas the summary functions lists the 12 plots but also gives how many observations are in each plot - so you can also see which plots had more/less observations.

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.

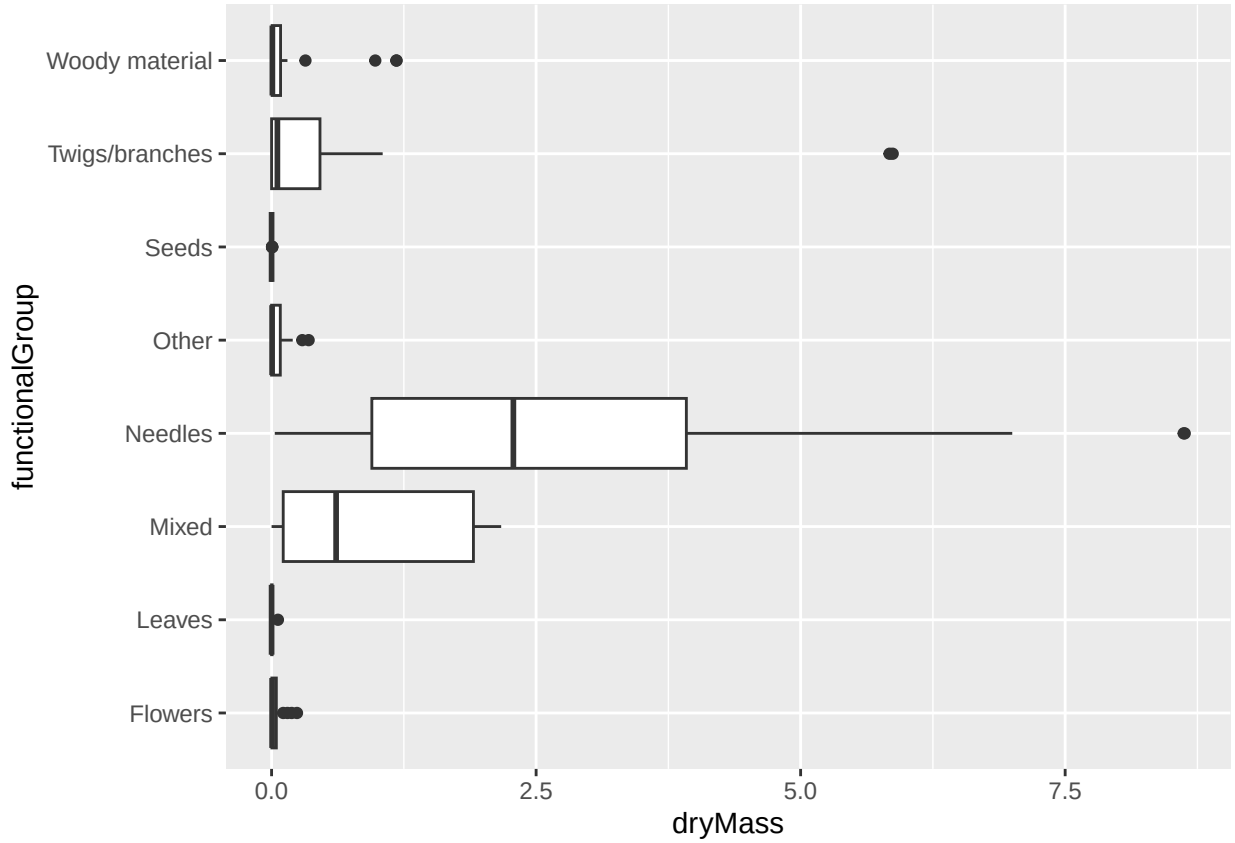
```
ggplot(Litter) +
  geom_bar(aes(x= functionalGroup))
```



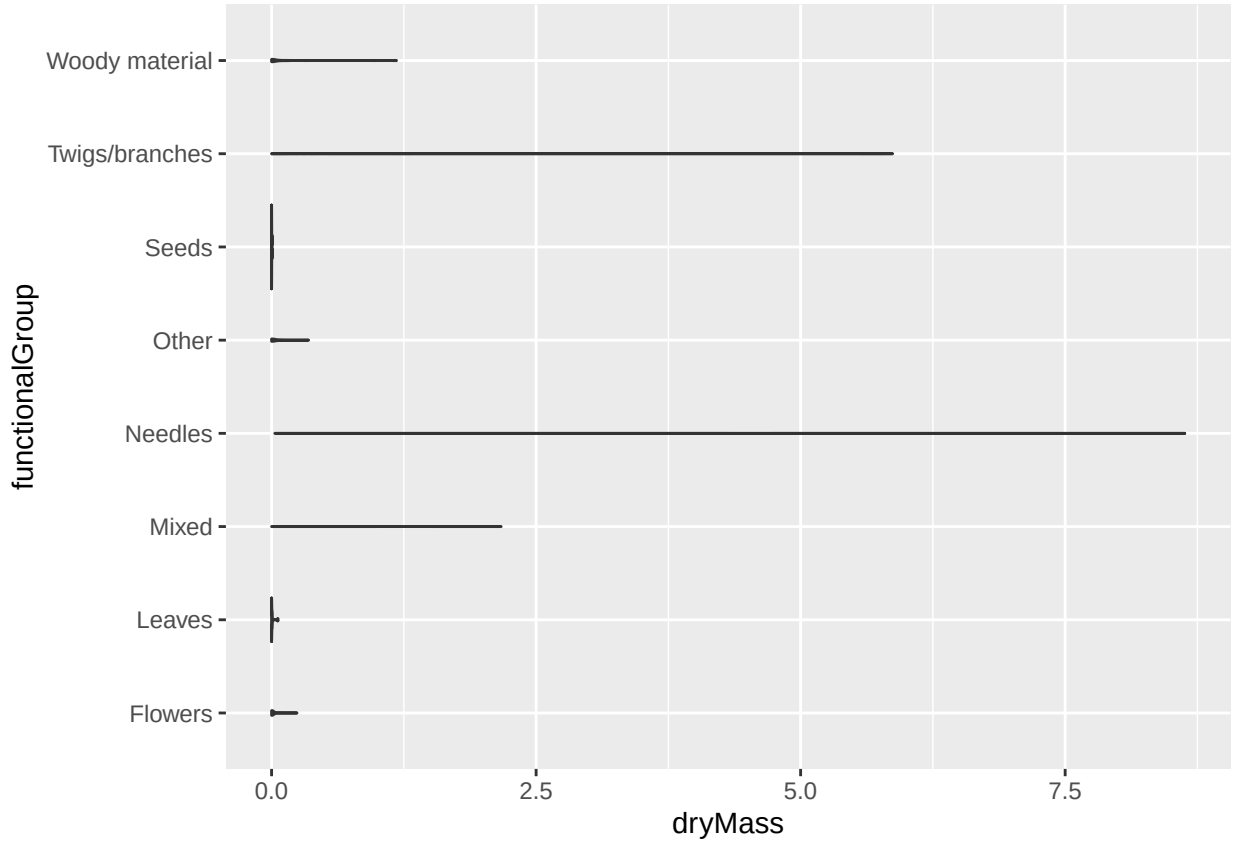

```
#Played around with bins and 50 gives a good visualization of the data.
```

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))
```




```
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: The boxplot allows to see outliers and also how is the distribution (quartiles, percentiles, etc) of the observations whereas the violin plot aggregated everything together

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

> Answer: Needles.