Assignment 3: Data Exploration

Sky Volz

Spring 2026

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. [NEW] 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 Canvas.
8. Initial here to acknowledge that you did not use AI in completing this assignment, except where expressly allowed: SV

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

TIP: If your code fails to knit, check: * That no `install.packages()` or `View()` commands exist in your code. * That you are not displaying the entire contents of a large dataframe in your code.

Set up your R session

1. Load necessary packages (tidyverse, here), check your current working directory and import 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: * Use the `here()` package in specifying the paths to your datasets * Include the appropriate subcommand to read in character based columns as factors

```r
#loading packages
library(tidyverse); library(here)

#checking working directory
here()
```

```
## [1] "/home/guest/EDE_Spring2026_personal"
```

```r
#read in data, stringsAsFactors used to make sure character based columns show up as factors
Neonics.data <- read.csv(
  file = here("Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv"),
  stringsAsFactors = TRUE)

Litter.data <- read.csv(
  file = 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. (AI is allowed here, but put answers in your own words.)

   Answer: Neonicotoids are insecticides used to get rid of pests in agriculture, but through dust, leaching, plant uptake through pollen, and spray drift these insecticides end up in the surrounding environment and spreads to other non-target insects. It is persistent and water soluble which makes it especially harmful. This insecticide is lethal and can also change behaviors for many differen insects, and pollinators are especially threatened which makes this a major environmnetla issue. This dataset measures how much insects have been exposed to the chemicals and how they have been effected by it. This is important to study because it threatens pollinators and other beneficial insects, which could cause food web collapse and other major issues. Also, if insects are being effected then this means the insecticide is polluting soil and waterways which is how the insects would be exposed to it.

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. (AI is allowed here, but put answers in your own words.)

   Answer: Measuring litter on the forest floor can provide important information about the health of a forest. Litter is vital for soil health and nutrient cycling, because as litter decomposes on the forest floor it releases nutrients into the soil. Woody debris is also important for carbon sequestration and the amount of debris present over time could signal decomposition rates and how long it takes for carbon to be sequestered. The amount of debris present is also important for moisture retention and providing habitat for organisms. Lastly, the amount of litter present could impact fire risk.

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. Litter is collected in elevated 0.5m2 PVC traps. Litter is defined as a material that is dropped from the forest canopy and has a butt end diameter <2cm and a length of <50cm. #2. Fine wood debris is collected in ground traps which are sampld once per year. Fine wood debris is defined as material that is dropped from the forest canopy and has a butt end diameter of <2cm and a length >50cm. #3. One litter trap pair, one elevated trap and one ground trap, is deployed for every 400 m2 plot area, resulting in 1-4 trap pairs per plot.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
neonics_dimension <- dim(Neonics.data) #using the dimension command

print(neonics_dimension)
```

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

6. Using the `summary` function on the "Effect" column, determine the most common effects that are studied. [Tip: The `sort()` command is useful for listing the values in order of magnitude...]

```
sort(summary(Neonics.data$Effect)) #sorted the summary stats for the most common effects
```

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

Question: Which two effects stand out as the most studied? Can you guess why these effects might specifically be of interest? > Answer: Population and mortality are the most studied. This is most likely because population changes and death rates to show immediate and long-term impacts of insectides on important insect species.

7. Using the `summary` function, determine the six most commonly studied species in the dataset (common name).[TIP: Explore the help on the `summary()` function, in particular the `maxsum` argument...]

```
#maxsum = 6 sets the amount of levels to be shown for that factor that is being summarized, which is co
summary(Neonics.data$Species.Common.Name, maxsum = 6)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee               (Other)
##                   152                   140                  3196
```

Question: What do these species have in common? Why might they be of interest over other insects? > Answer: All of these species are pollinators, which make them of extreme interest because without pollinators there would be a mass die-off of plants. Flowering plants require pollination from pollinators to reproduce, so without pollinators these plants would die.

8. The `Conc.1..Author` column, which lists the concentration of the neonicitoid dose, should include numeric values. 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...]

```
class(Neonics.data$Conc.1..Author.) # checking the class
```

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

```
view(Neonics.data$Conc.1..Author.) #looking at the column in the dataset
```

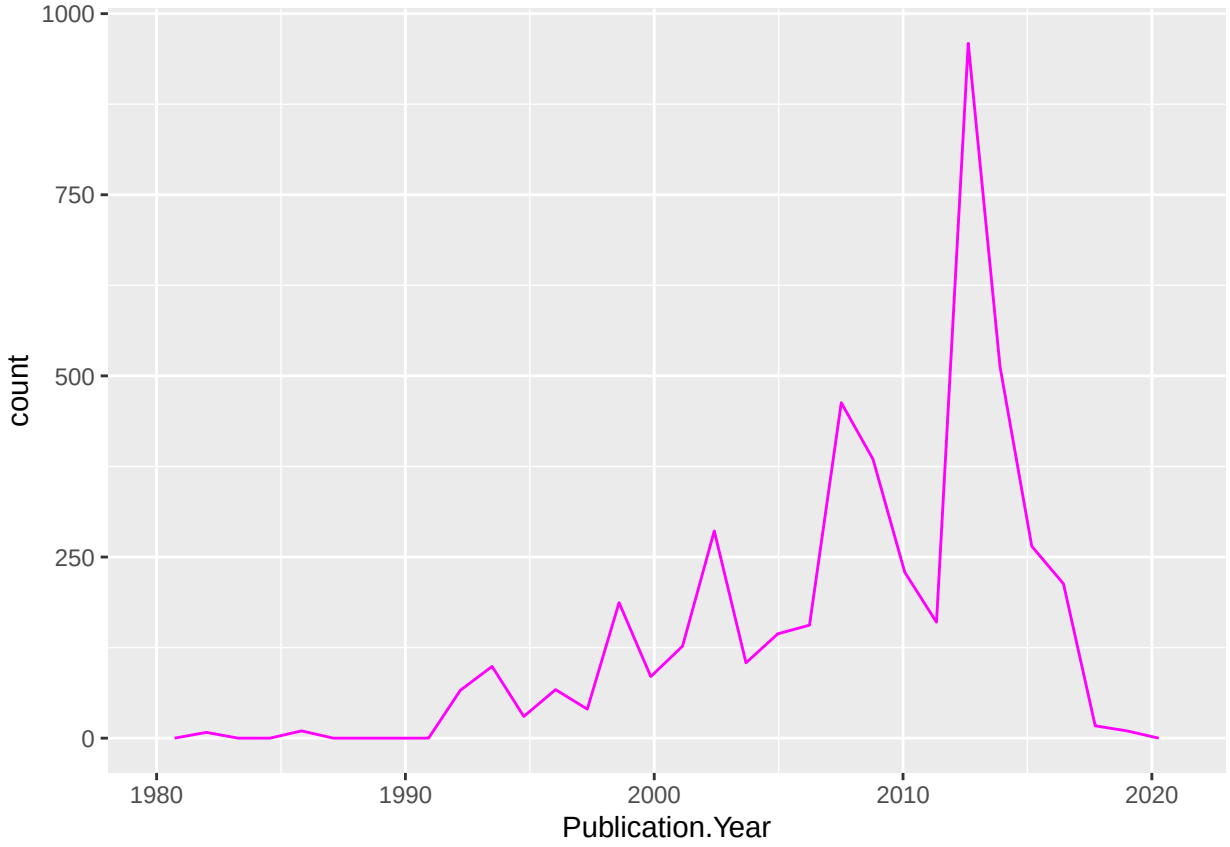
> Answer: Looking at the values in the column in the dataset, some of the values are "NR" or there is a /, which is why it will not be classified as numeric as there are non-numbers. If would be listed as a character class, but I changed this to be a factor when I read in the data.

Explore your data graphically (Neonics)

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

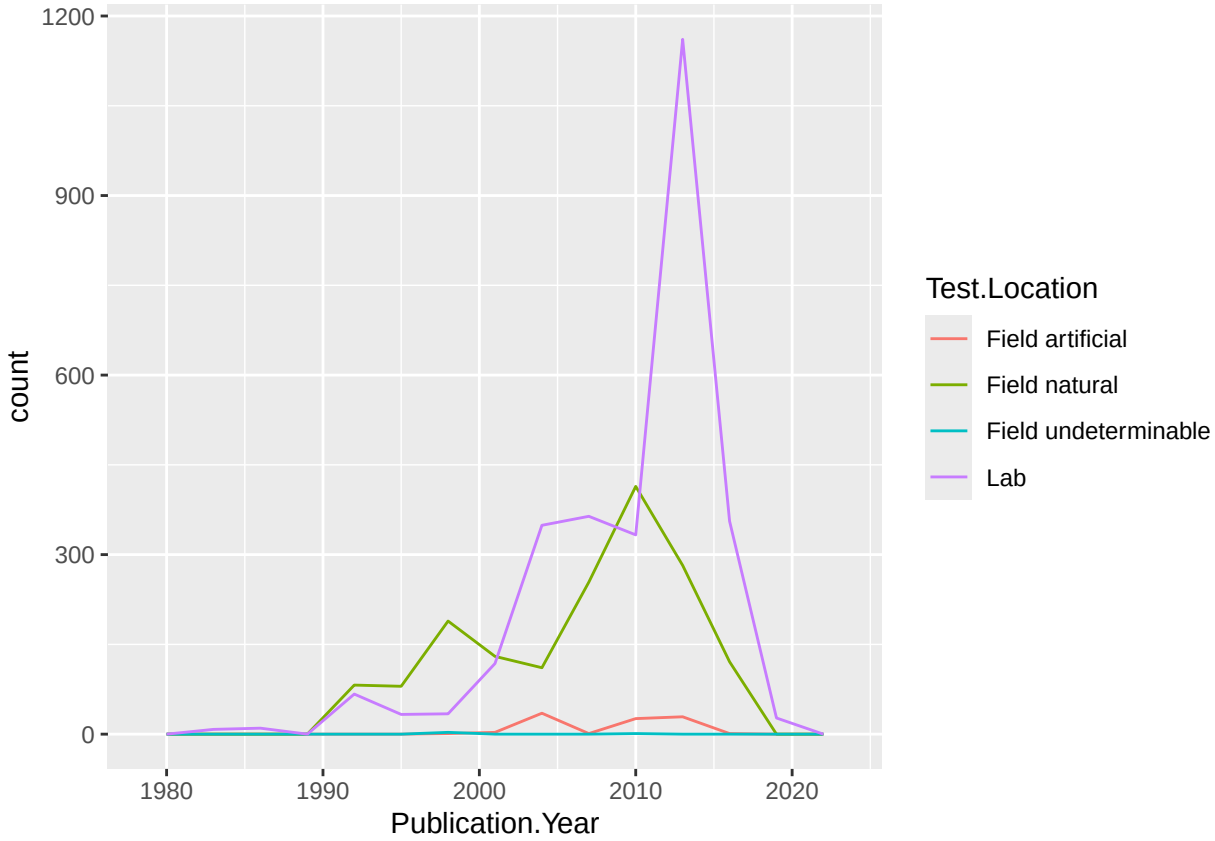
```
ggplot(Neonics.data, aes(x=Publication.Year)) + #set x axis as number of studies by publication year
  geom_freqpoly(color = "magenta") #turned the line pink
```

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



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

```
  #set x axis as number of studies by publication year, assigning a color to test locations
ggplot(Neonics.data, aes(x=Publication.Year, color = Test.Location)) +
  geom_freqpoly(binwidth=3) #set small binwdith to show more detail
```

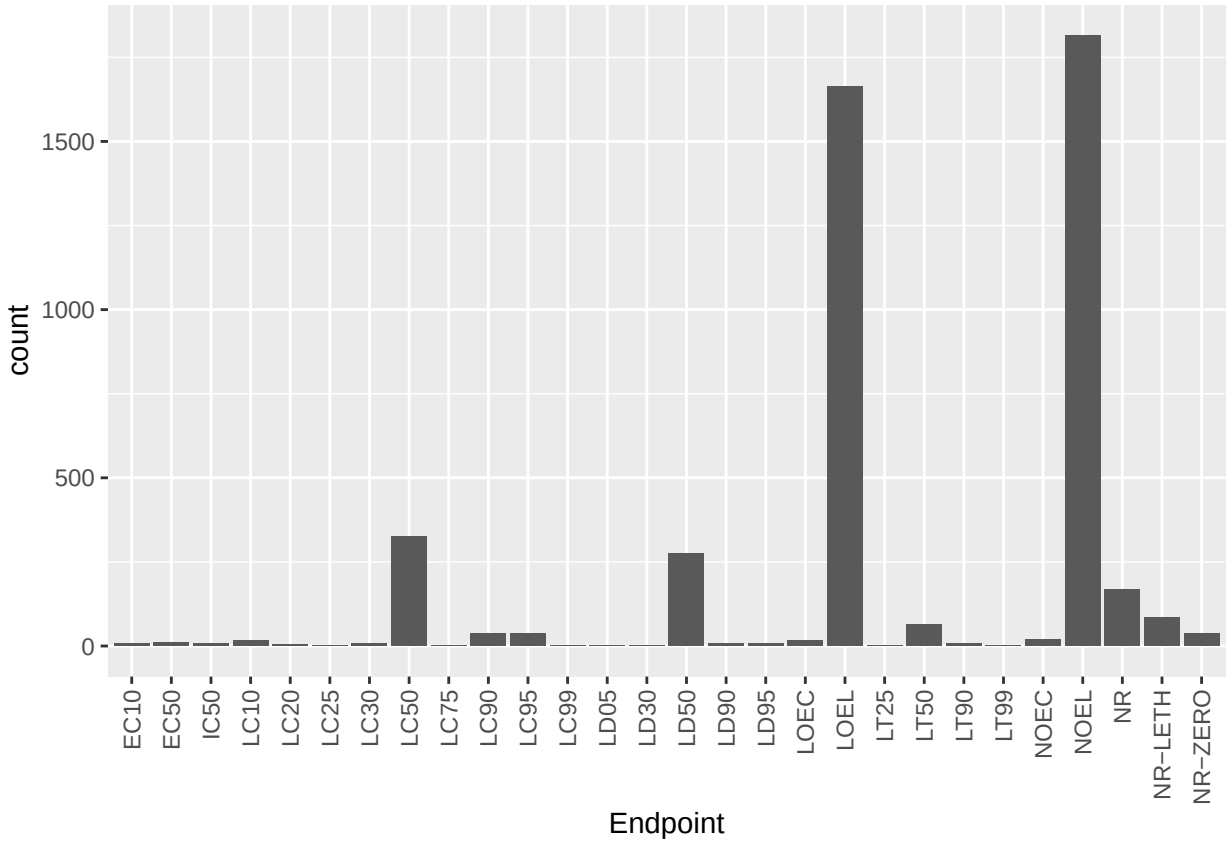


Interpret this graph. What are the most common test locations, and do they differ over time? > Answer: Prior to 2000, the most common test locations were from natural fields, but soon after 2000 the most common test location was in a lab. This change happened slowly at first, but the amount of tests being conducted in labs increased dramatically soon after 2010, while the number of tests occuring in natural fields steadily decreased.

11. Create a bar graph of Endpoint counts.

[**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.data, aes(x=Endpoint)) +
  geom_bar() +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) #to format labels
```



What are the two most common end points, and how are they defined? Consult the ECOTOX_CodeAppendix (p.721) for more information. > Answer: The two most common end points are NOEL and LOEL. LOEL is terrestrial data, and is defined the lowest observable effect level: lowest dose (concentration) producing effects that were significantly different from responses of controls (p. 722). NOEL is also terrestrial data and is defined as no observable effect level: the highest does (concentration) producing effects not significantly different from responses of controls according to the author's reported statistical test (p.723).

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.data$collectDate) #looking at class of collectDate
```

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

```
Litter.data$collectDate <- ymd(Litter.data$collectDate) #used lubridate to convert the dates to a date

class(Litter.data$collectDate) #checking that it worked
```

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

```
unique(Litter.data$collectDate) #found that data was sampled on Aug 2 and Aug 30 in 2018
```

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

13. Using the `unique` function, list the different `plotIDs` sampled at Niwot Ridge.

```
unique(Litter.data$plotID) #listing unique plotIDs
```

```
##  [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
```

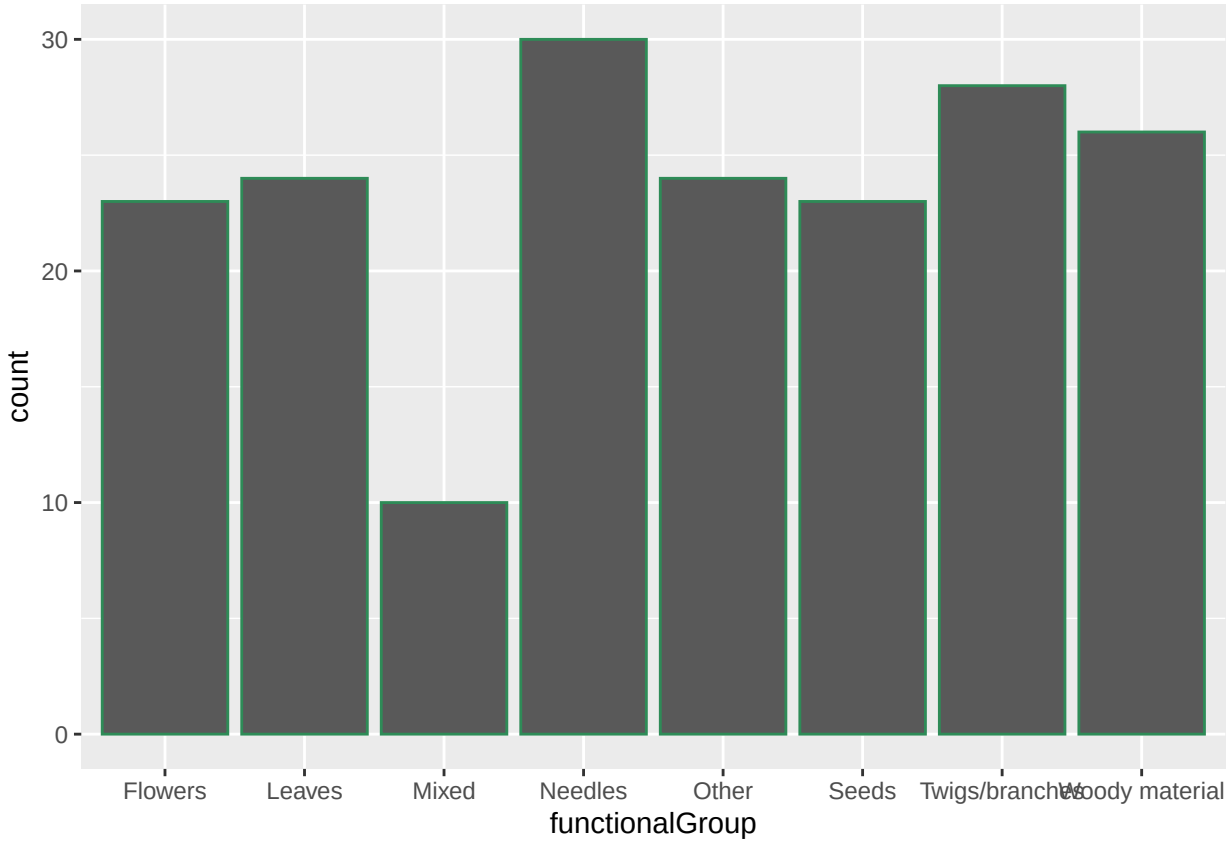
```
summary(Litter.data$plotID) #checking to see what summary would show to compare with unique
```

```
## 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
```

How is the information obtained from `unique` different from that obtained from `summary`? > Answer: The information from unique shows how many different plot IDs were sampled, and summary shows how many samples were take at each plot ID.

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.

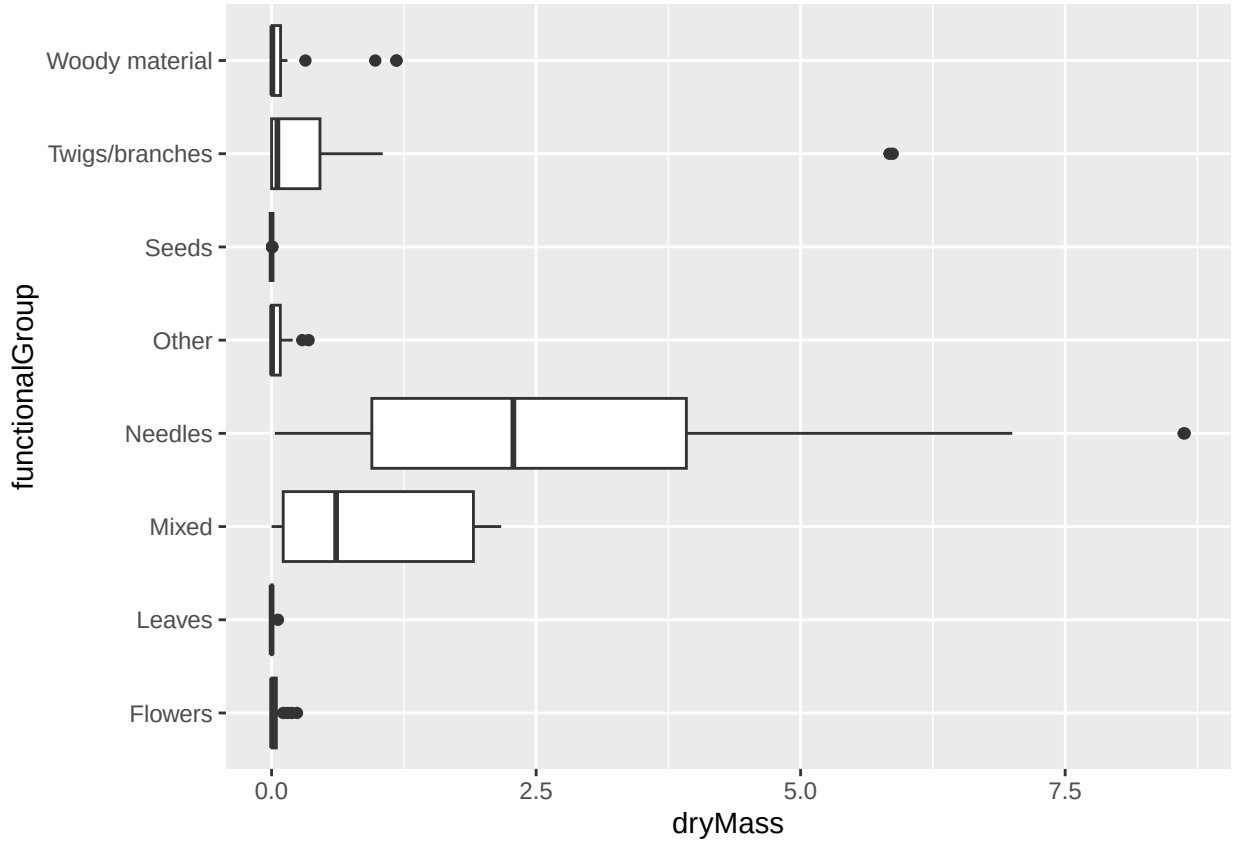
```
ggplot(Litter.data, aes(x=functionalGroup)) +
  geom_bar(color="seagreen")
```



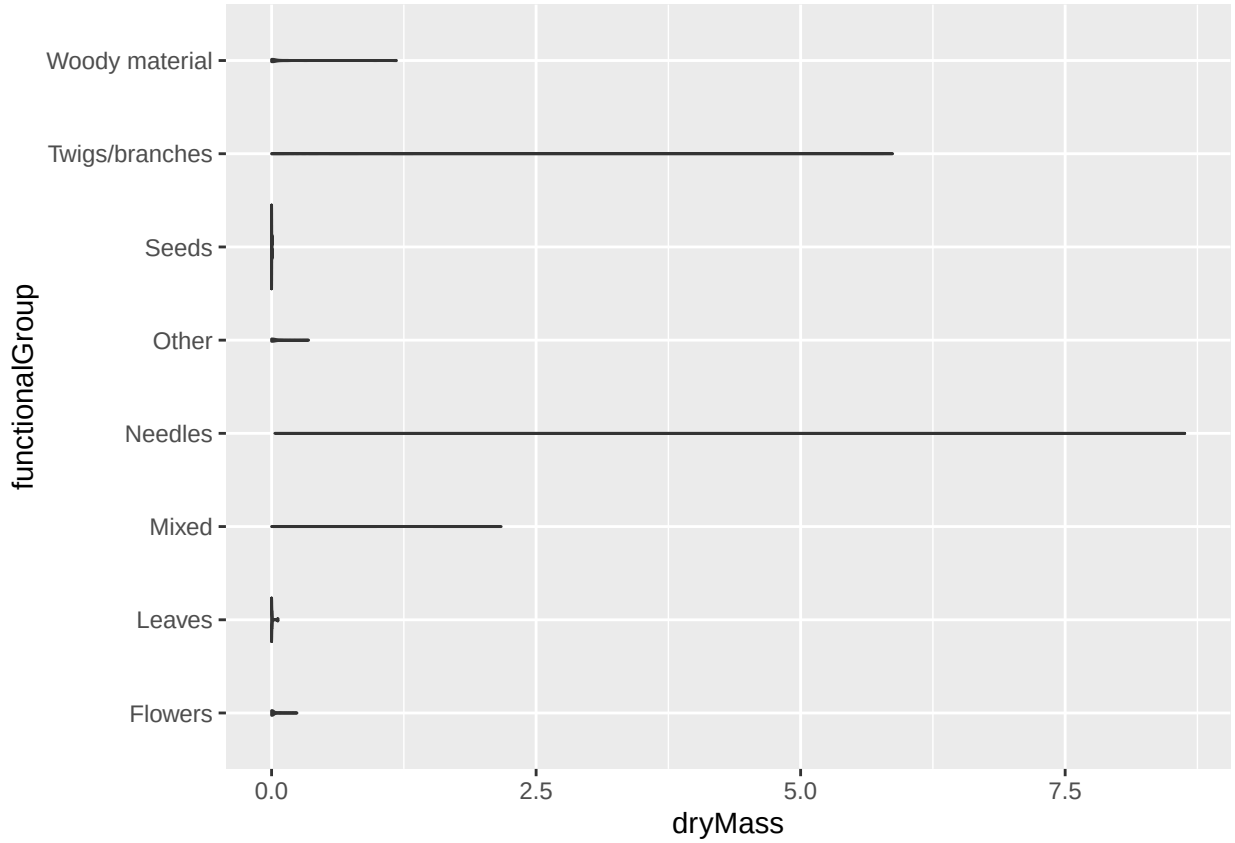

```
#used geom_bar to create a bar graph and set the x-axis to show groups of litter types
#outlined it in green
```

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

```
#create boxplot, setting the x-axis as dry mass of litter and the y-axis as type of litter
ggplot(Litter.data, aes(x=dryMass, y=functionalGroup)) +
  geom_boxplot()
```




```
#create violin plot, setting the x-axis as dry mass of litter and the y-axis as type of litter
ggplot(Litter.data, aes(x=dryMass, y=functionalGroup)) +
  geom_violin()
```



Why is the boxplot a more effective visualization option than the violin plot in this case? > Answer: The boxplot is a more effective visualization because it shows the presence of outliers in the data, where the violin plot does not. For example, the violin plot shows that the dryMass of needles is greater than 7.5, however the box plot shows that on average this mass is around 2.5 with an outlier value of almost 9. This is also true for twigs/branches, which has a large outlier that skews the data in the violin plot.

What type(s) of litter tend to have the highest biomass at these sites? > Answer: The box plot shows that the sites tend to have the highest biomass of needles and "mixed" litter, because these have the highest median value compared to the other types of litter.