Assignment 3: Data Exploration

Melissa Merritt

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()  #This allows me to ensure that my working directory is what I want it to be.
```

```
## [1] "/home/guest/Documents/EDA-Fall2022"
```

```
library(tidyverse)  #the library() function allows me to add the tidyverse package.

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)

# The commands above allow me to import the data, label the data, and ensure
# that it is imported # as factors.
```

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 would be interested in the ecotoxicology of neonicotinoids because the levels of neonicotinoids in the environment can, directly and indirectly, affect many different species.

Specifically, it is essential to study the impact of the neonicotinoid on insects because scientists have observed a decline in many insect populations due to these pesticides. Pollinators, like bees, are significantly affected by the use of neonicotinoids, and we have already seen how a loss of pollinators can affect an ecosystem. Neonicotinoids are banned in the European Union because of their adverse environmental effects, so we would be interested in studying the ecotoxicology of neonicotinoids on insects because it might lead to new policies in the US.

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: One of the main reasons we would be interested in studying litter and woody debris on the ground in forests in the west is to determine wildfire risk. Observing the fuel load on a forest floor can allow you to determine a landscape's ignition potential and flammability. This can provide information on how a fire might spread on the ground, and if ladder fuels are present in the forest, it can be a great determiner of how a fire might spread into the canopy. Another reason to study litter and woody debris on the forest floor is to determine a forest's decomposition rates, which can inform forest health.

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 dry weight of litter and woody debris were collected from litter traps organized by functional plant type. 2. The spatial sampling design was defined by sampling at terrestrial NEON sites that contained woody vegetation >2m tall. This sampling occurred in randomly selected tower plot locations. 3. The temporal sampling data defined the target sampling frequency as based on whether the forest was deciduous or evergreen. The ground traps are sampled once per year.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

```
# the data dimensions (dim() function) gives you the number of rows and columns
# in the data.
```

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
```

```
# I am able to find the information for a single column using $.
```

Answer: The most common effects studied are population and mortality. The Mortality effect would be of specific interest because it would inform us of correlated or causal relationships between neonicotinoids' use and insects' death. This is one of the most pressing interests because if it is causing insect death, that can have a more significant impact on the ecosystem. The population would also be of specific interest because it would tell us the number of species that are still in the ecosystem, and that would give us insight into the longer-term effects of the pesticides.

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, 7)
```

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

```
# I chose to calculate the seven most commonly studies species because the
# other category does not give us much information.
```

Answer: The six most commonly studied species of insects are all pollinators. As I mentioned, pollinators are inordinately impacted by pesticides because of their relationships with plants. Pollinators are also a key species in the ecosystem because many plants rely on them to reproduce, and many other species of insects and animals rely on these plants for survival.

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.)
```

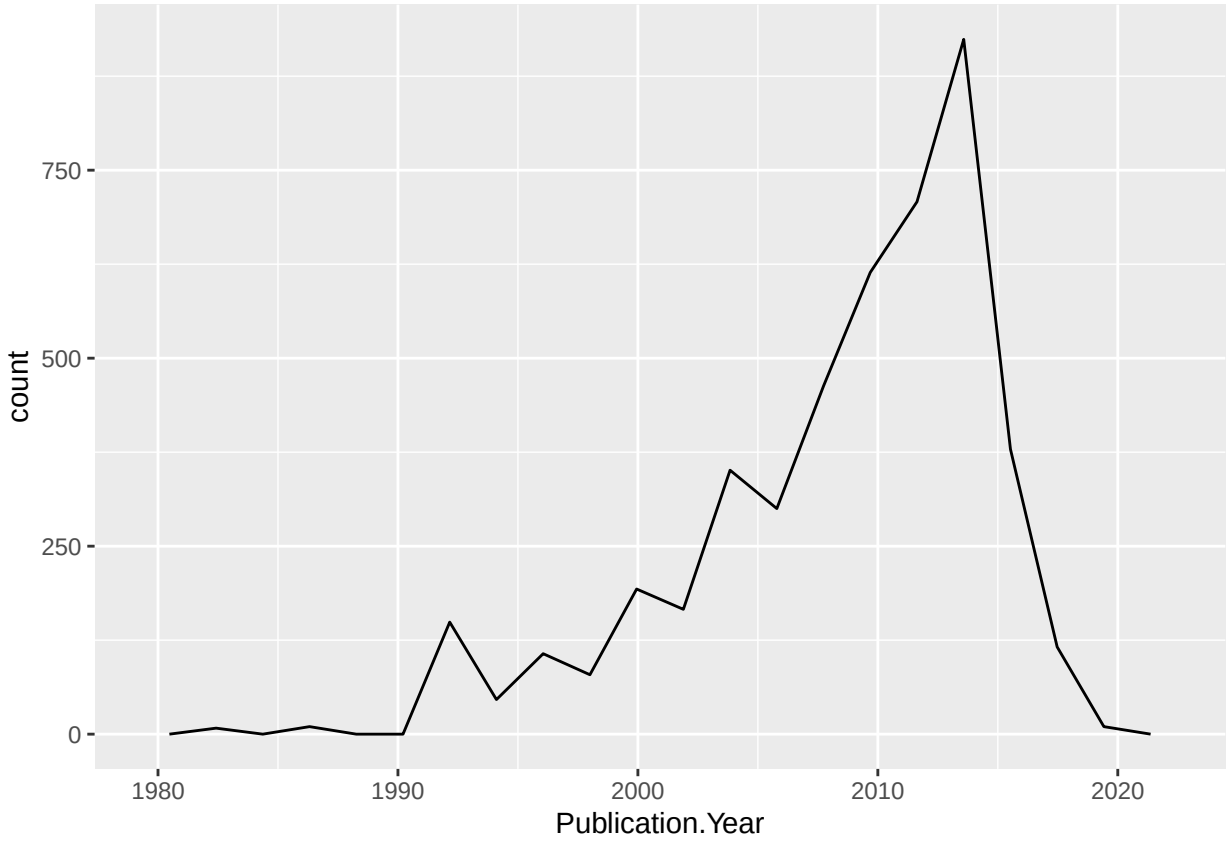
```
## [1] "factor"
```

Answer: The class of Conc.1..Author is defined as a factor because we set the values as factors when we imported the dataset. Factor is a discrete value, while numeric is not.

Explore your data graphically (Neonics)

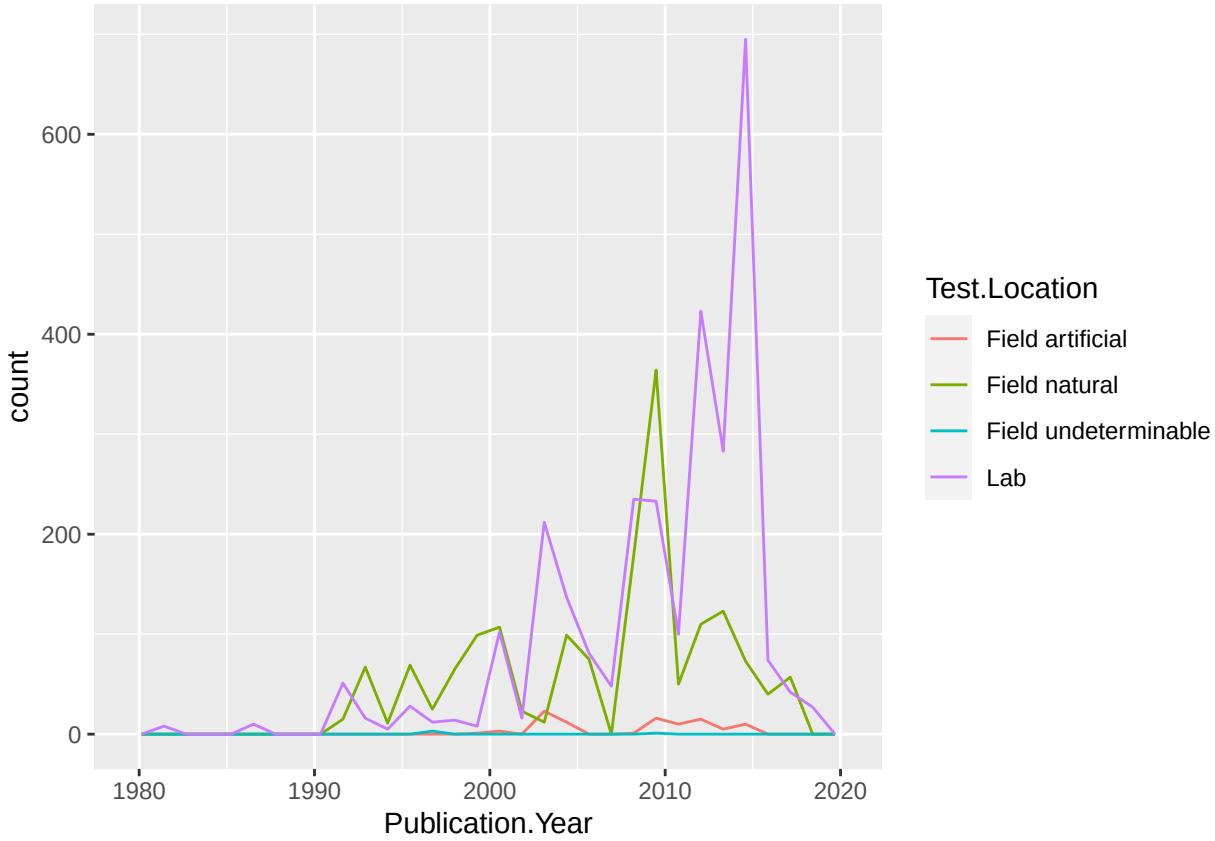
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics) + geom_freqpoly(aes(x = Publication.Year), bins = 20)
```



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),
    bins = 30)
```

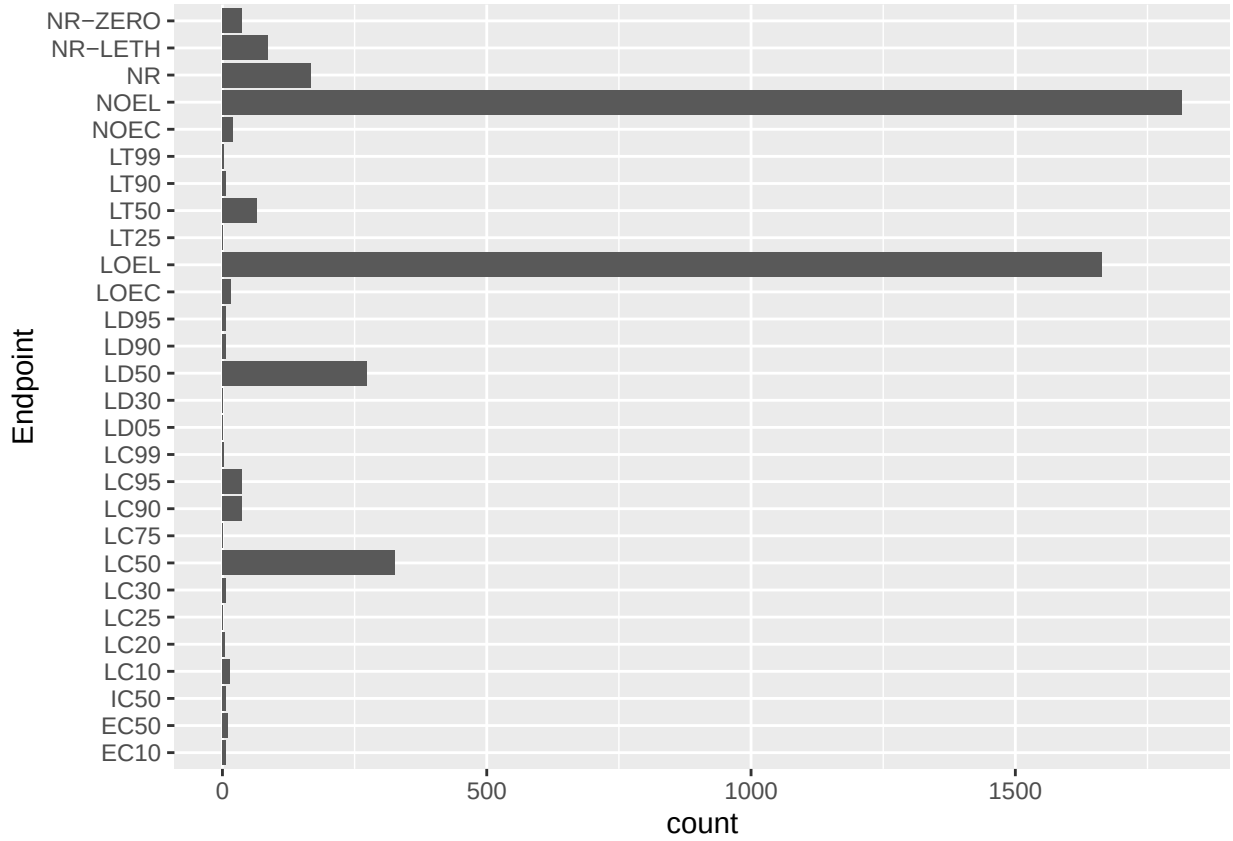


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are the Lab and in the field (natural). At the beginning (~1992 - 2000), the field (natural) testing locations were more common. In about 2004, Lab testing started to increase, and it peaked in 2014. Field (natural) testing locations mainly were less than lab testing except for a few spikes, the most considerable spike being in 2009. Getting closer to 2020, both of the testing site locations decrease substantially.

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) + geom_bar(aes(x = Endpoint)) + coord_flip()
```




```
# since the axis titles were not visible, I used the coord_flip() function to
# make the bar chart horizontal.
```

Answer: The two most common endpoints are NOEL and LOEL. NOEL stands for No-observable-effect-level, and this means the highest dose producing effects not significantly different from the responses of controls. LOEL stands for Lowest-observable-effect-level, which means that it is the lowest dose producing effects that were significantly different from the control responses.

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"
```

```
# First time I ran the code to find that the class for collectDate was a
# factor, so I used the as.Date function to change it to a date, and then
# checked again to see that the class was now 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)
```

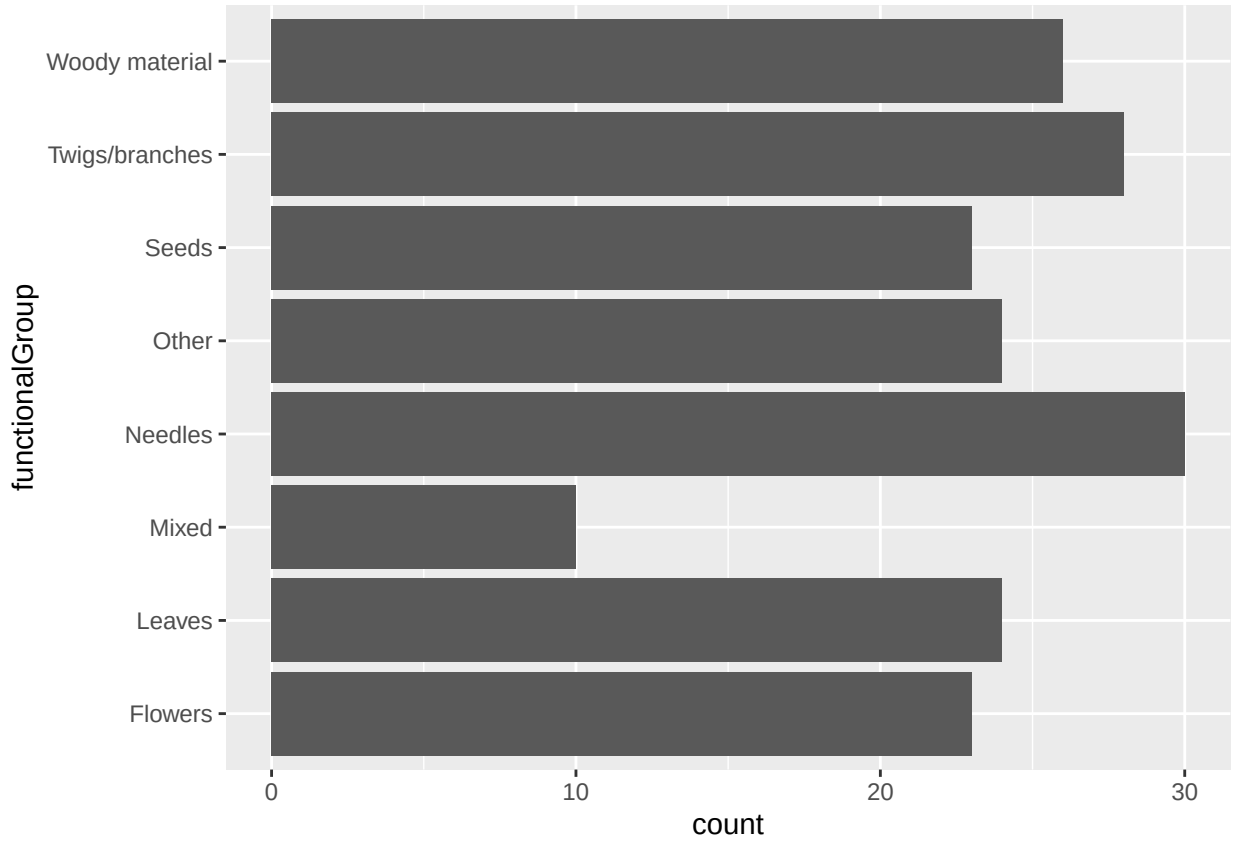
```
##  [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

```
# Using the unique and count functions I was able to count the amount of plots
# sampled at Niwot Ridge.
```

Answer: There were 12 plots sampled at Niwot Ridge. The unique function allows you to see how many plots were sample because it counts each factor once, and does not count reoccurances of the factor. If we were to use the summary function with the same information it would tell us how many samples were in 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.

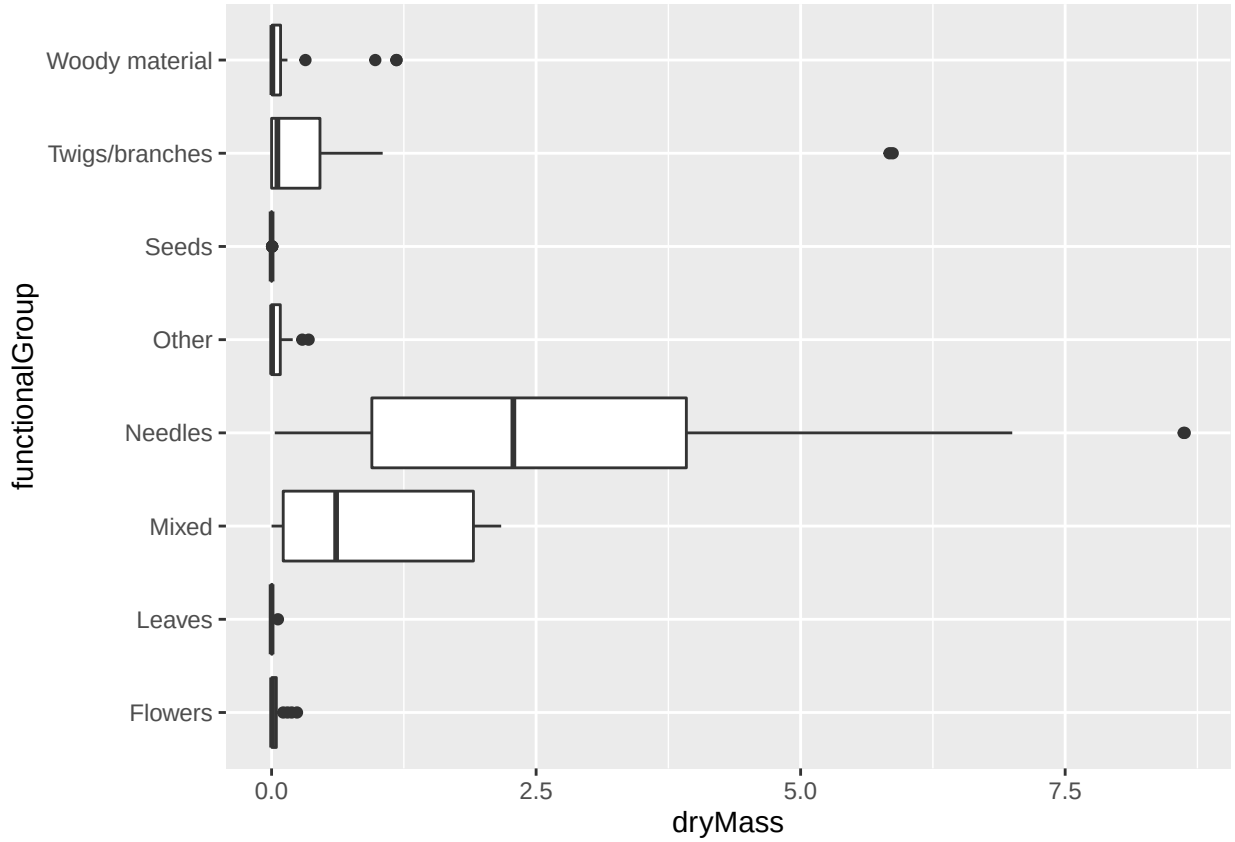
```
ggplot(Litter) + geom_bar(aes(x = functionalGroup)) + coord_flip()
```



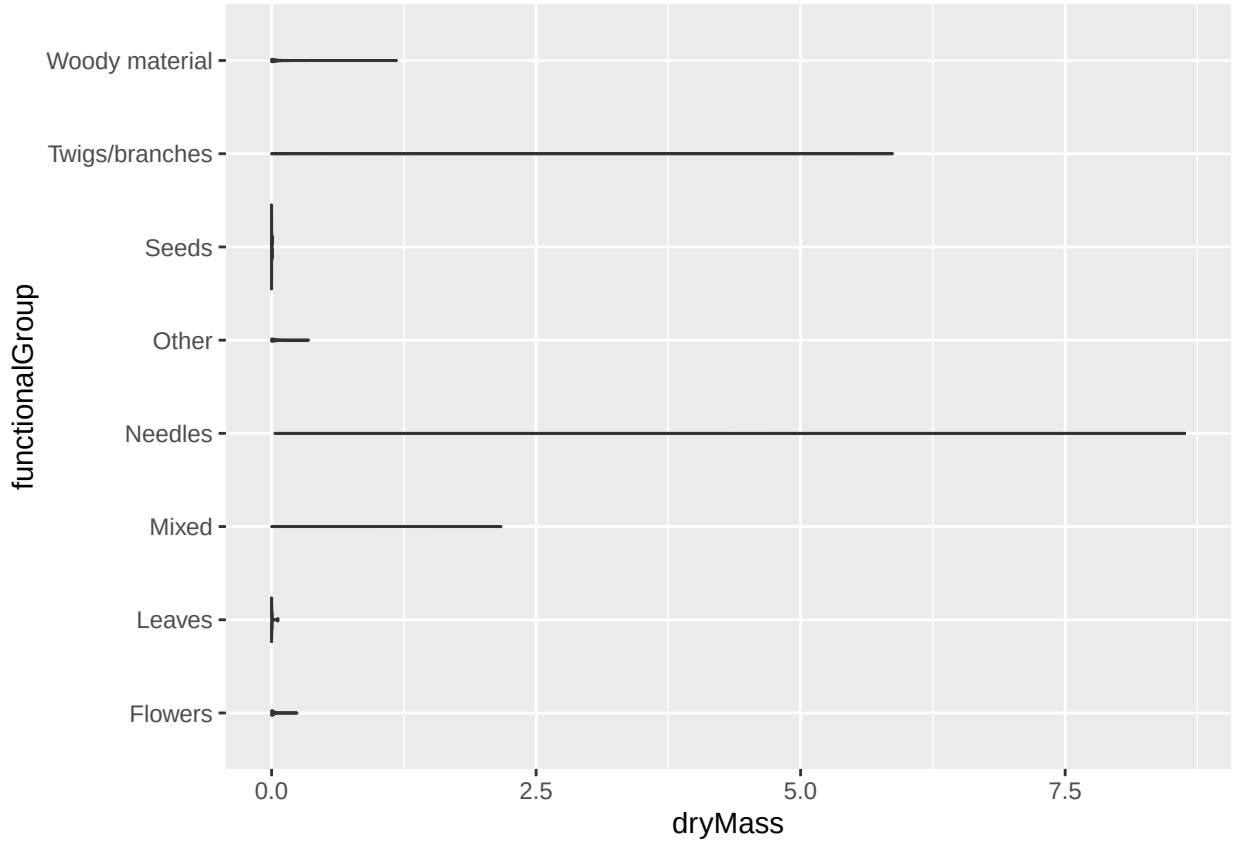

```
# Using ggplot() allows me to graph a bar chart with geom_bar(). I defined my
# variable as the functionalGroup, and once again flipped the bar chart, so all
# the axis labels could be seen.
```

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))
```




```
# Using ggplot() with the two different graph types I was able to define the x
# and y axes to create two different types of graphs.
```

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

> Answer: The boxplot is a better visualization option than the violin plot in this case because it highlights which of the data points are outliers. In addition, the boxplot allows you to see the actual distribution of the dryMass, including visuals of the means, the interquartile range, and the rest of the distribution. The violin plot, in this case, only shows you the entire distribution from smallest to largest, but not where the majority of the data is.

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

> Answer: The three types of litter with the highest biomass at these sites are needles, mixed and twigs/branches.