Assignment 3: Data Exploration

Jonathan Gilman

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.

```
#load packages
library(tidyverse)
library(lubridate)
library(here)
library(tinytex)

update.packages("rmarkdown")
#tinytex::install_tinytex()  # Ensure TinyTeX is installed and updated

#check current working directory
here()
```

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

```
#upload datasets

neonics <- read.csv(
  file = here('Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv'),
  stringsAsFactors = T
)
litter <- read.csv(
  file = here('Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv'),
  stringsAsFactors = T
)
```

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: My family back in Texas hobby beekeeps so growing up I was very interested in studying honeybees and entomology in general. So I know that neonicotinoids are very controversial as they can haved drastic negative effects on beneficial pollinator commmunities. However, they can be very effective at killing pests, so understanding how they affect different insects, and how those interactions might be changing over time is of great interest.

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: My research background is heavily focused on nutrient cycling, so for me, studying the litter would be interesting in terms of quantifying N,C, and P cycling in these ecosytems. This could be related to denitrification in these systems. It could also be related to the build-up of organic matter and the implications that has for wildfires.

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. In sites with $< 50\%$ cover of woody vegeta on, sites with heterogeneously distributed, patchy, vegeta on, trap placement is targeted such that only areas beneath qualifying vegeta on are considered for trap placement. 2. Ground traps are sampled once per year. 3. Sampling for this product occurs only in tower plots.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```r
# look at dimensions of neonics: # of rows, # of columns
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? [Tip: The `sort()` command is useful for listing the values in order of magnitude…]

```r
#summary of "Effect" column
summary_effect <- summary(neonics$Effect)

#sort by magnitude
summary_effect_sorted <- sort(summary_effect)
```

> Answer: The most common effects that are studied are population, mortality, and behavior. These effects might be specifically of interest because they are the most general and relevant. Especially for the testing of an insecticide, it makes sense that population and mortality of insects are more common response variables than things like hormones and histology – which are much more specialized.

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…]

```r
#summary of "Species.Common.Name" column
summary_species <- summary(neonics$Species.Common.Name)

#sort by magnitude
summary_species_sorted <- sort(summary_species)
```

> Answer: The six most commonly studied species in the dataset are the Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, and Italian Honeybee. All of these species are pollinators. They might be of interest over other insects because they more directly come into contact with insecticides that are applied. For example, a honeybee visits many many plants in a single day, but an ant might not come into contact with a plant on an average day.

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…]

```r
#find out the class of "Conc.1..Author."
class(neonics$Conc.1..Author.)
```

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

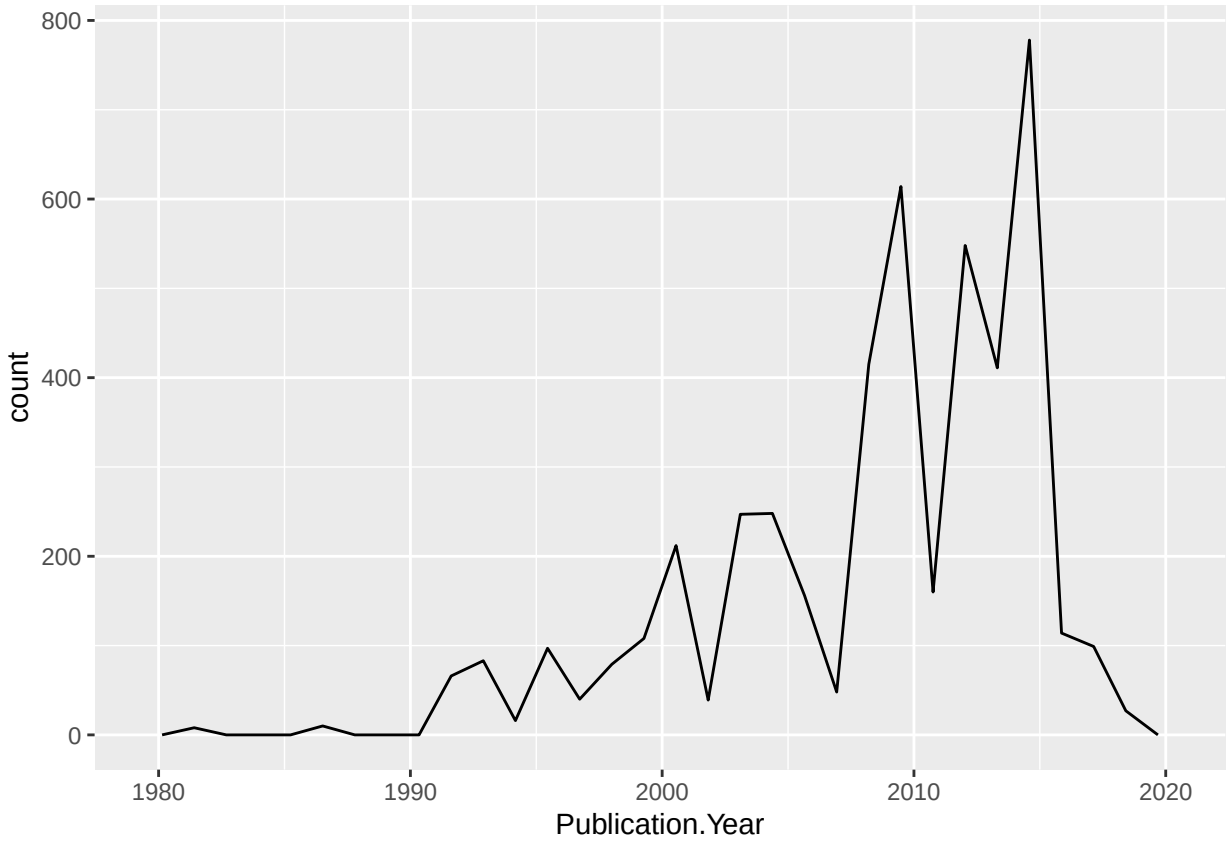
> Answer: The class for this column is a factor. It is not numeric because there are non-numeric characters within this column (/).

Explore your data graphically (Neonics)

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

```
# plot of number of studies conducted by publication year
studies_by_year <- ggplot(neonics, aes(x = Publication.Year))+
  geom_freqpoly()
studies_by_year
```

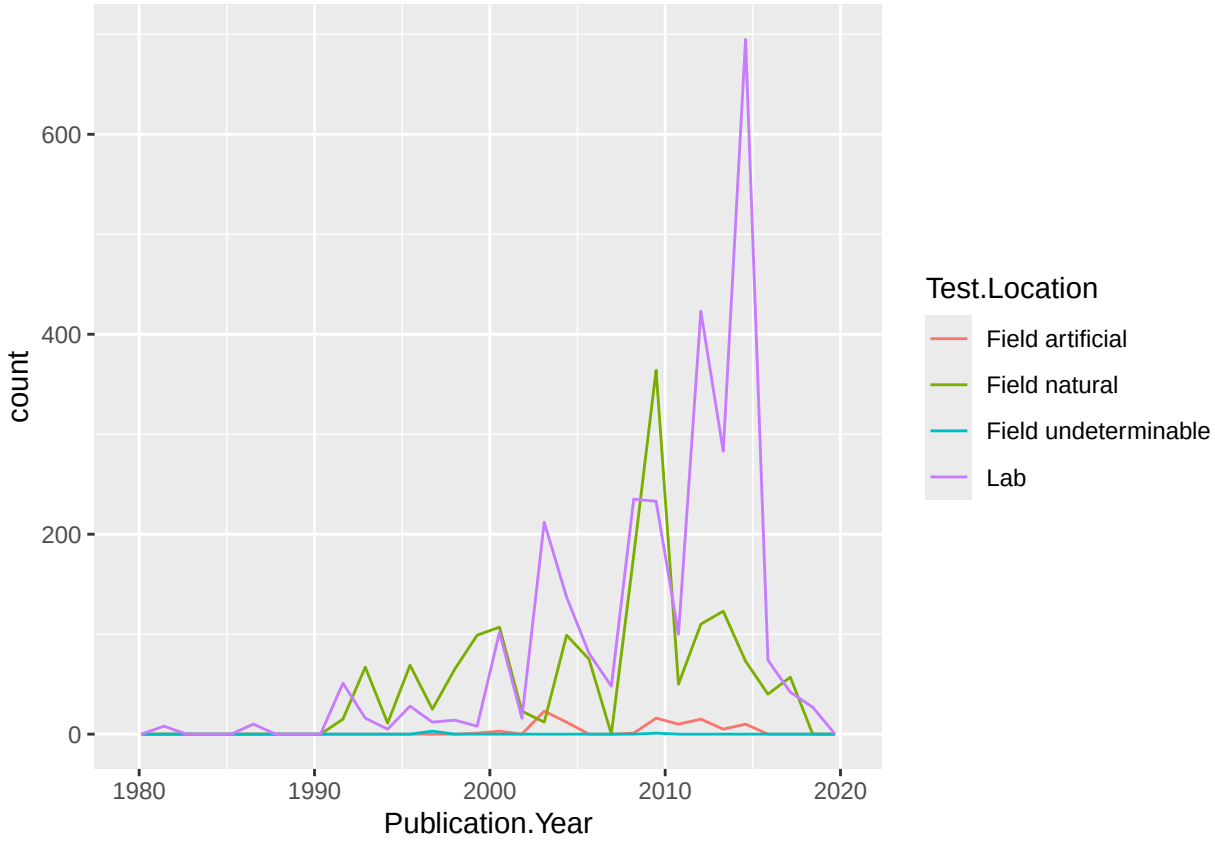
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
# plot of number of studies conducted by publication year with color
studies_by_year_color <- ggplot(neonics, aes(x = Publication.Year, color = Test.Location))+
  geom_freqpoly()
studies_by_year_color
```

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



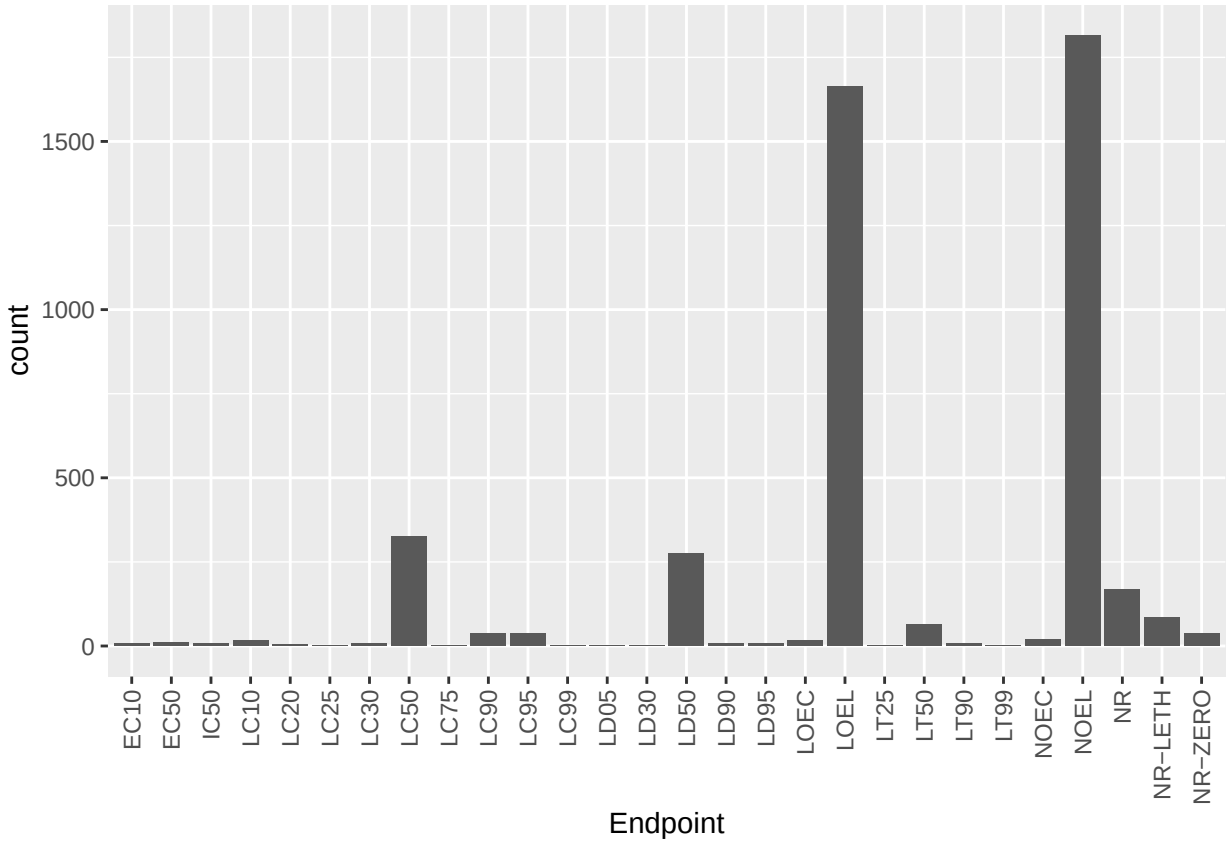
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations seem to be 'lab' and 'field natural'. All test locations seem to increase dramatically since 1990. So possibly this is when the effects of neonicotinoids started to come into question. It also seems like around 2009 'field natural' experiments reached a peak and then have declined since then. And as they declined, lab experiments greatly increased. One theory is that 'field natural' experiments were deemed unethical and have been replaced by more lab experiments.

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…]

```
# make a bar graph of Endpoint counts
endpoint_counts <- ggplot(neonics, aes(x = Endpoint)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))  # Rotate x-axis labels
endpoint_counts
```



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

```
# determine the class of collectDate
class(litter$collectDate)
```

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

```
# factor

# change collectDate class to a date
litter$collectDate <- as.Date(litter$collectDate, format = "%Y-%m-%d")

# determine new class
class(litter$collectDate)
```

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

```
#find out when litter was sampled in August 2018
august_2018_dates <- unique(litter$collectDate[litter$collectDate >= "2018-08-01" & litter$collectDate
august_2018_dates
```

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

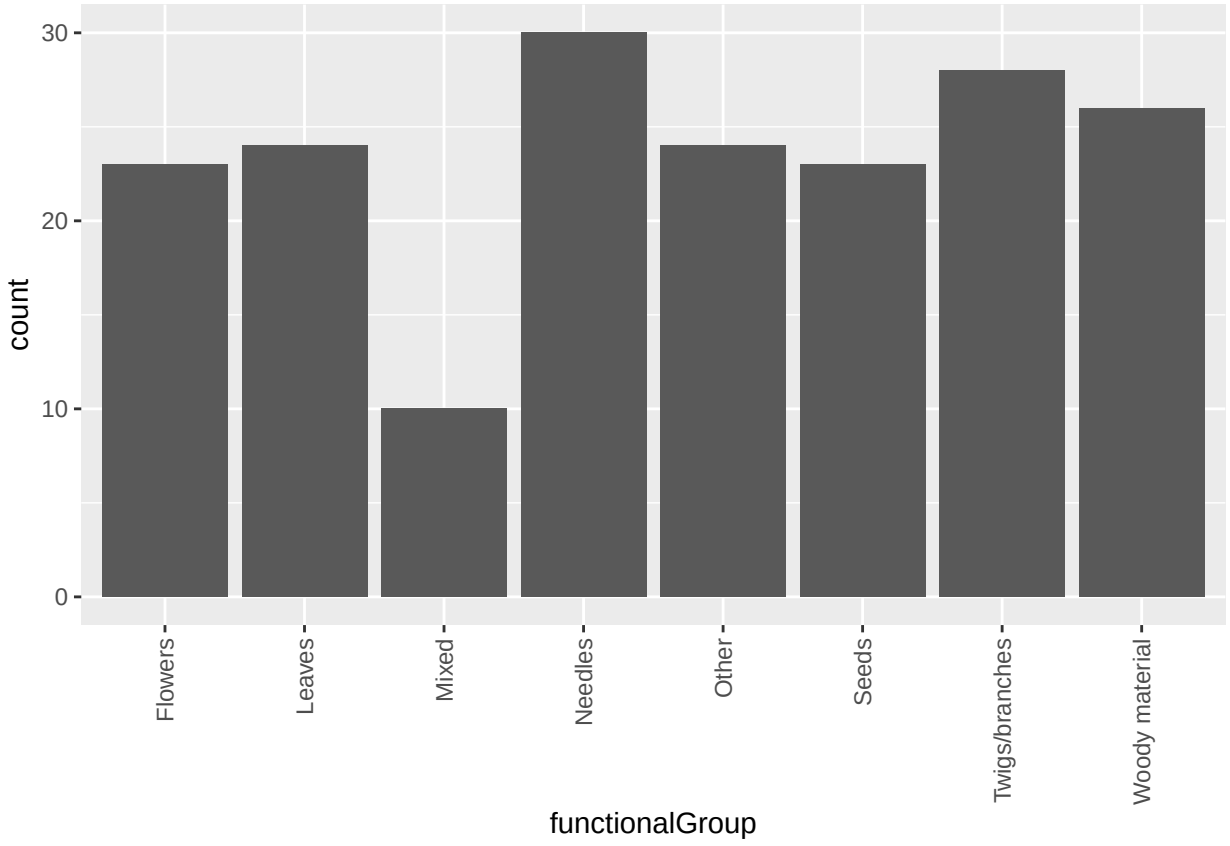
13. Using the `unique` function, determine how many different plots were sampled at Niwot Ridge. How is the information obtained from `unique` different from that obtained from `summary`?

```
# count unique plot IDs
unique_plots <- unique(litter$plotID)
number_of_unique_plots <- length(unique_plots)
#12
```

Answer: There were 12 different plots sampled at Niwot Ridge. This information obtained from 'unique' is different from that obtained from 'summary' because 'summary' would provide a count of each unique plot ID, but not the count of unique plot IDs.

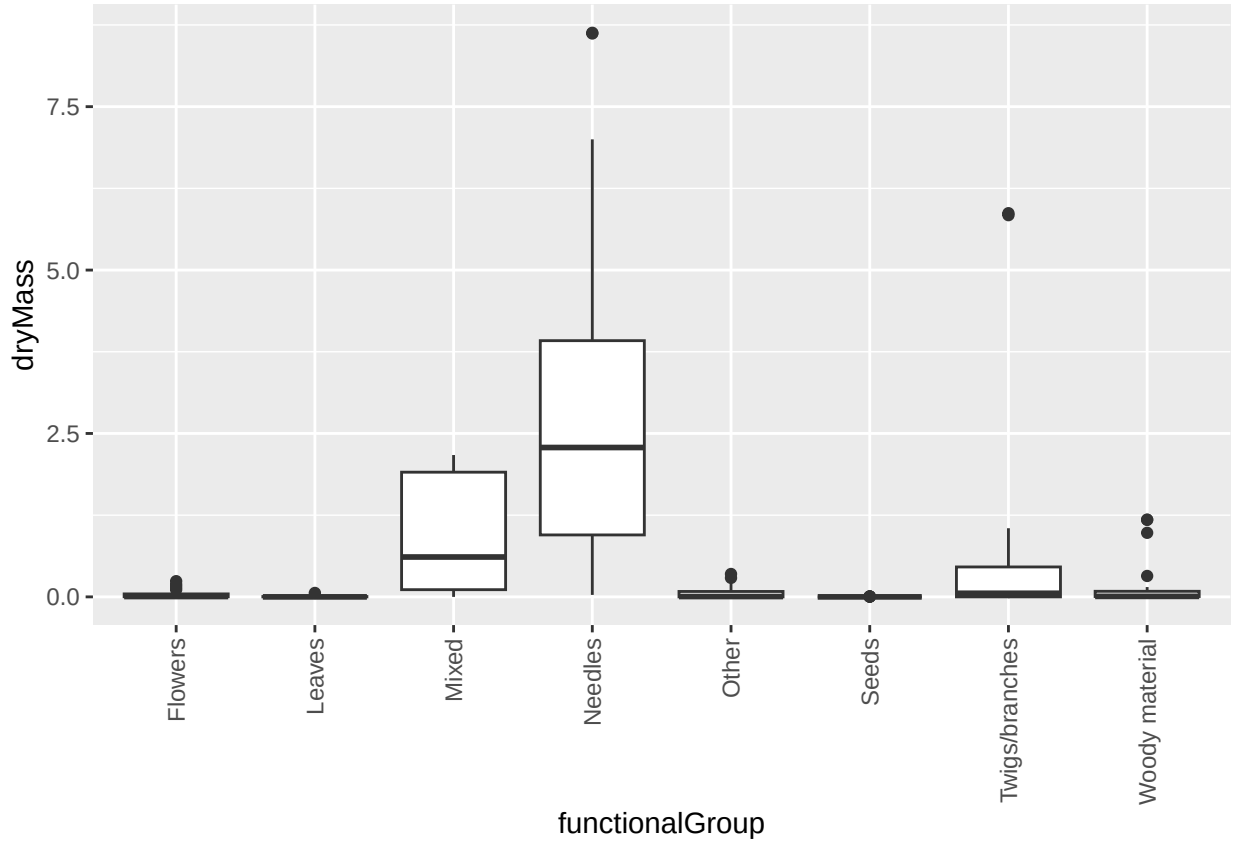
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
# bar graph of functionalGroup counts
functional_group_counts <- ggplot(litter, aes(x = functionalGroup)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
functional_group_counts
```

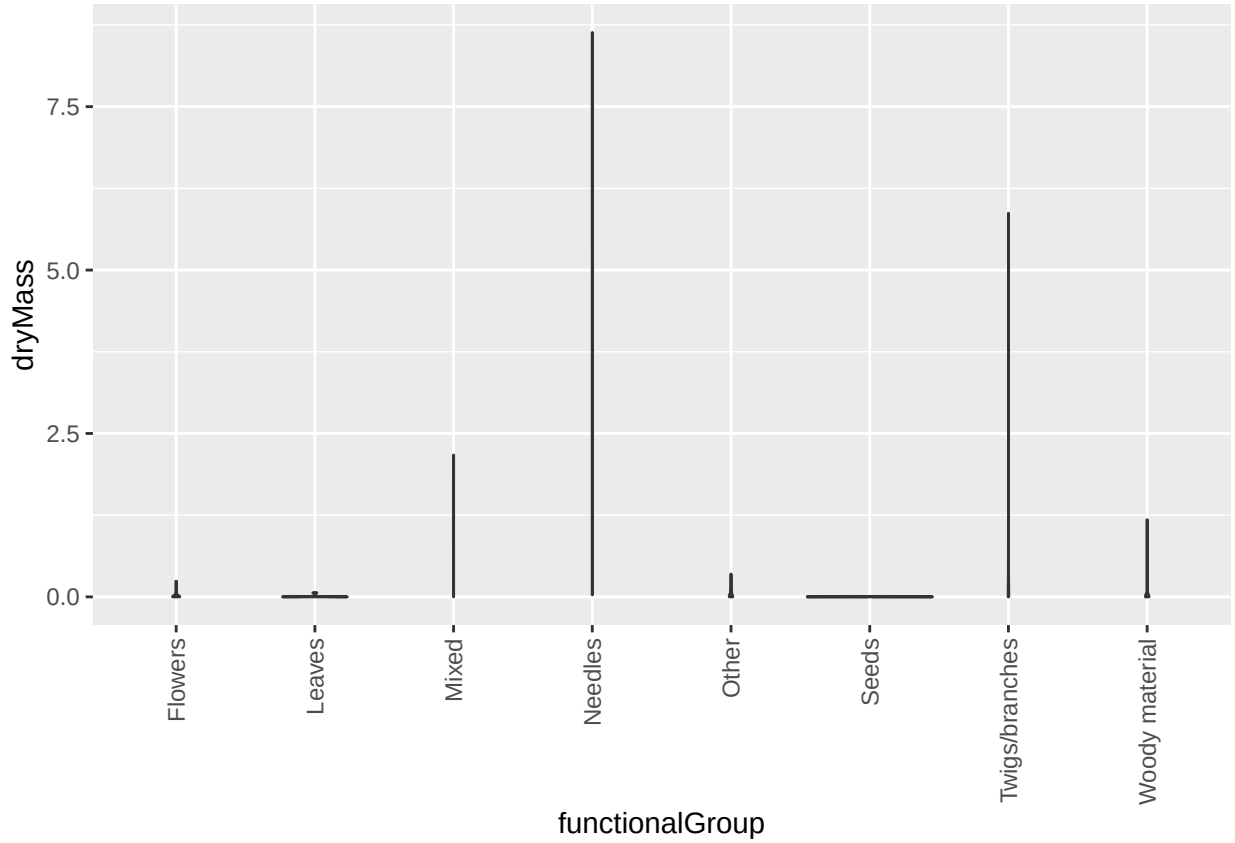


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#boxplot of dryMass by functionalGroup
boxplot_dryMass <- ggplot(litter, aes(x = functionalGroup, y = dryMass)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
boxplot_dryMass
```




```
#violin plot of dryMass by functionalGroup
violin_dryMass <- ggplot(litter, aes(x = functionalGroup, y = dryMass)) +
  geom_violin() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
violin_dryMass
```



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 clearly shows the distribution of dryMass by functionalGroup with means, quartiles, and potential outliers. The violin plot, on the other hand, is not very effective here possibly because we have a limited amount of data and they don't have much variation.

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

> Answer: "Needles" and "mixed" litter tend to have the highest biomass at these sites.