Sriya Gottumukkula_A03_DataExploration.Rmd

Sriya gottumukkula

Spring 2025

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.

```
# Loading packages (tidyverse, here)
library(tidyverse)
library(here)

# Checking the working directory

getwd()
```

```
## [1] "/home/guest/R course/EDA_Spring2025"
```

```
here()
```

```
## [1] "/home/guest/R course/EDA_Spring2025"
```

```
# Importing datasets
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.

```
#view(neonics.data)
```

Answer:

Studying the ecotoxicology of neonicotinoids is important because these insecticides, while effective against pests, also harm non-target insects like bees. Research links neonicotinoids to reduced reproduction, impaired foraging, and increased mortality in pollinators, threatening biodiversity and food security. Their persistence in the environment can disrupt ecosystems, making it crucial to assess their risks for sustainable pest management.

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:

Studying litter and woody debris is important for nutrient cycling, carbon storage, and ecosystem health. Decomposing material enriches soil, supports plant growth, and provides habitat for fungi, insects, and wildlife. Understanding these processes helps assess forest stability and inform conservation efforts.

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 Traps – Litter and fine woody debris are collected using elevated and ground traps, designed to capture different sizes and types of organic material.

2. Spatial Sampling – Sampling takes place in designated plots, with trap placement varying based on vegetation density and site characteristics.

3. Temporal Sampling – Collection frequency differs across ecosystem types, with regular sampling intervals adjusted for seasonal changes and site conditions.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(neonics.data)
```

```
## [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...]

```
summary(neonics.data$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
```

Answer:

The most common effects studied in the dataset are population changes (1,803 occurrences), mortality (1,493), reproduction (197), behavior (360), and feeding behavior (255). These effects are significant because they highlight the ecological risks of neonicotinoid pesticides. High mortality and population declines can threaten biodiversity, particularly for pollinators like bees, which play a crucial role in agriculture and ecosystems. Reproductive effects indicate potential long-term declines, even at sublethal doses. Behavioral and feeding changes, such as disorientation and reduced foraging, can weaken insect populations and disrupt food chains. Studying these effects helps assess pesticide toxicity and informs descion making

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...]

```
summary(neonics.data$Species.Common.Name, maxsum = 5)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee               (Other)
##                   152                  3336
```

Answer:

The six most commonly studied species include honey bees, parasitic wasps, buff-tailed bumblebees, and Carniolan honey bees, with the remaining grouped under "Other." These species are key pollinators and natural pest controllers, making them crucial for agriculture and biodiversity. Their high study frequency is likely due to concerns over pesticide exposure, habitat loss, and population declines, which impact ecosystems and food security

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...]

```
class(neonics.data$Conc.1..Author.)
```

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

```
#view(neonics.data$Conc.1..Author.)
```

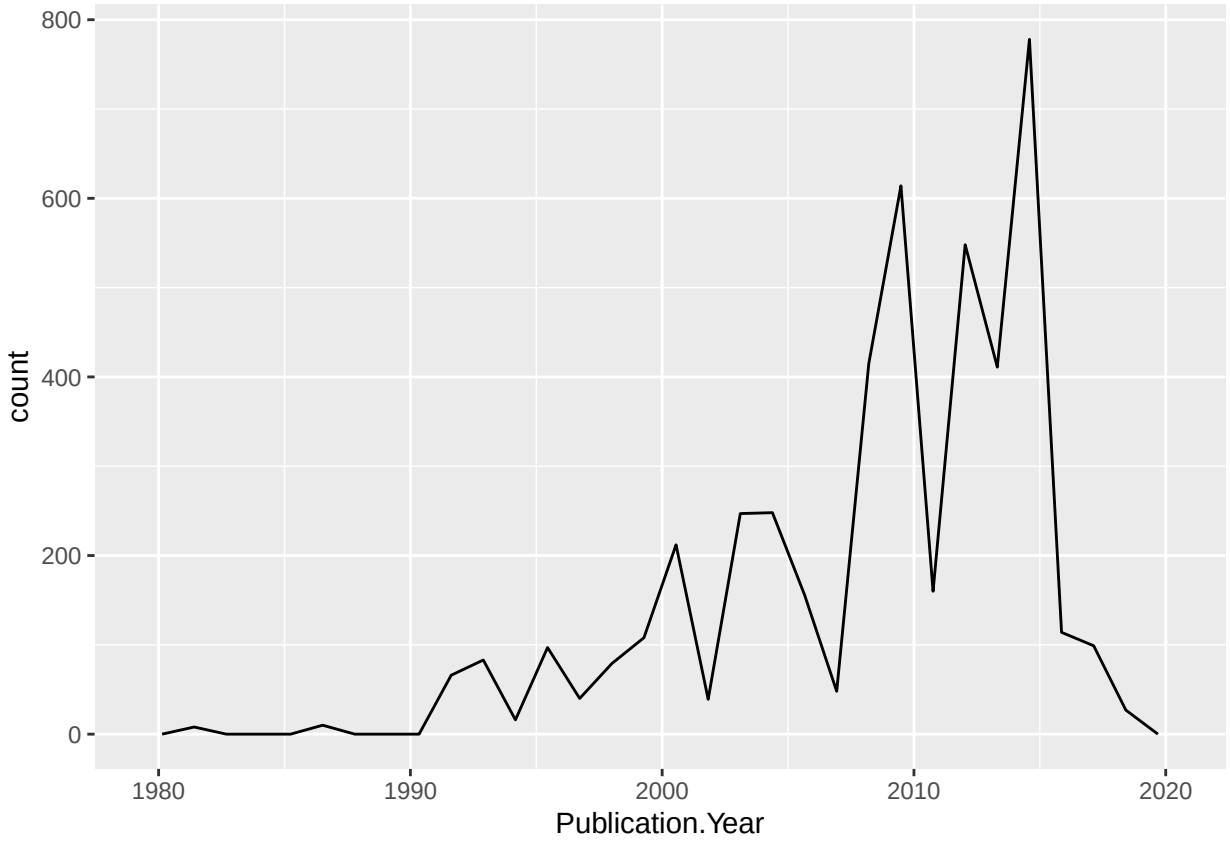
Answer:R is reading the column `Conc.1..Author.` as factor and not numerical because there are some non numeric values in te data set like NR thus it is reading as factor

Explore your data graphically (Neonics)

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

```
ggplot(neonics.data) +
  geom_freqpoly(aes(x = Publication.Year)
  )
```

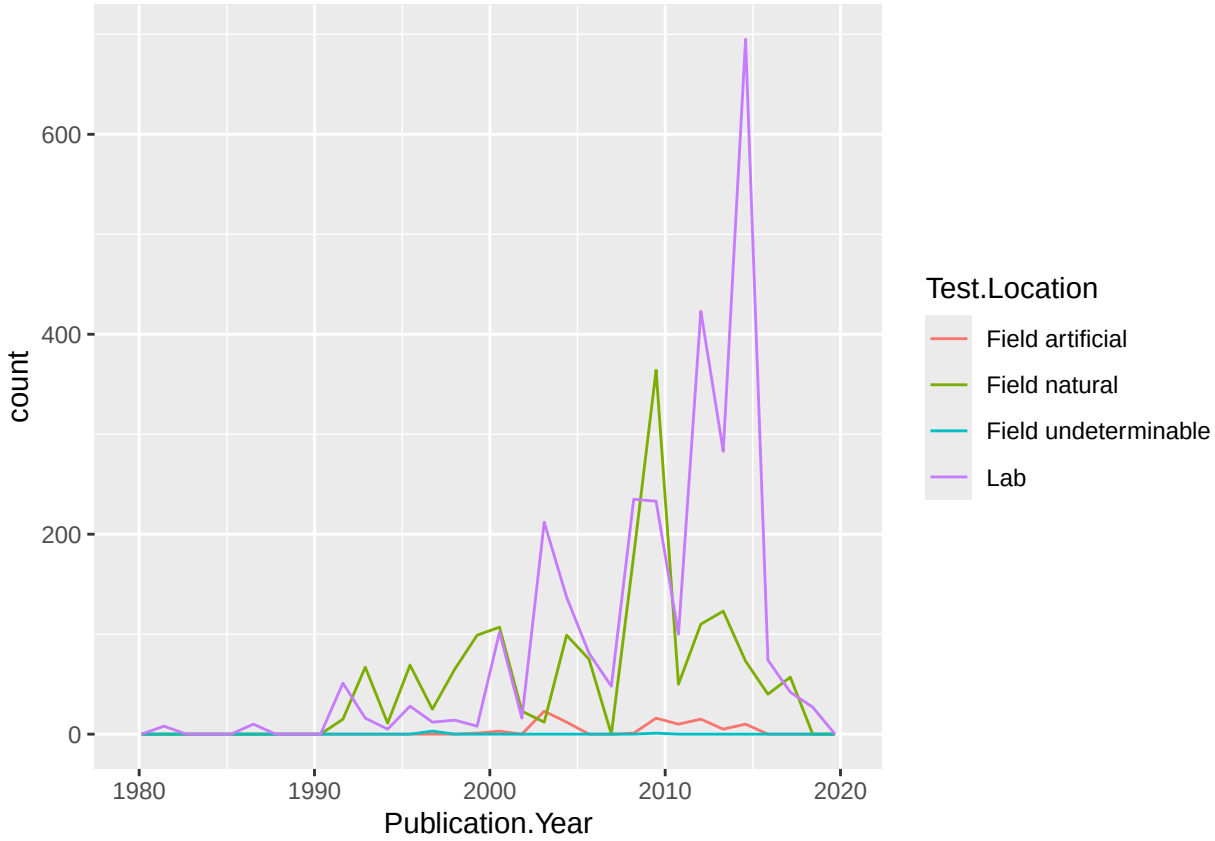
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```
ggplot(neonics.data) +
 geom_freqpoly(aes(x = Publication.Year, color = Test.Location))
```

```
## `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?

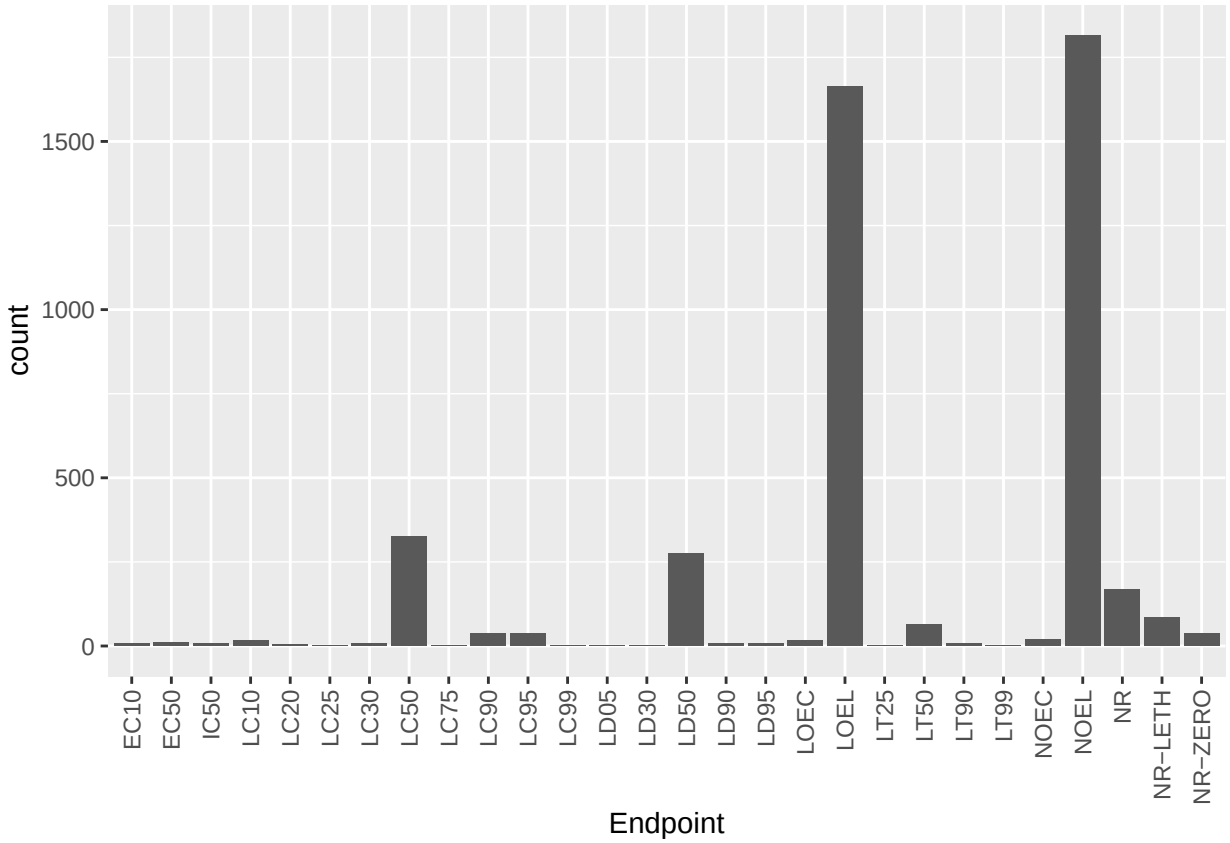
Answer:

The most common test locations are lab and they tend to be most common test locations especially during 2010 to 2016 but before and after this period they were not as common with others like feild natural also being one of the common test locations across.

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



Answer:

the two most common end points are and they are defined as:

1. NOEL - 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)
2. Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC)

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)
```

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

```
#view(litter.data$collectDate)

litter.data$collectDate <- as.Date(litter.data$collectDate, format = "%Y-%m-%d")

unique(litter.data$collectDate)
```

```
## [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`?

```
unique(litter.data$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
```

```
summary(litter.data$plotID)
```

```
## 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
```

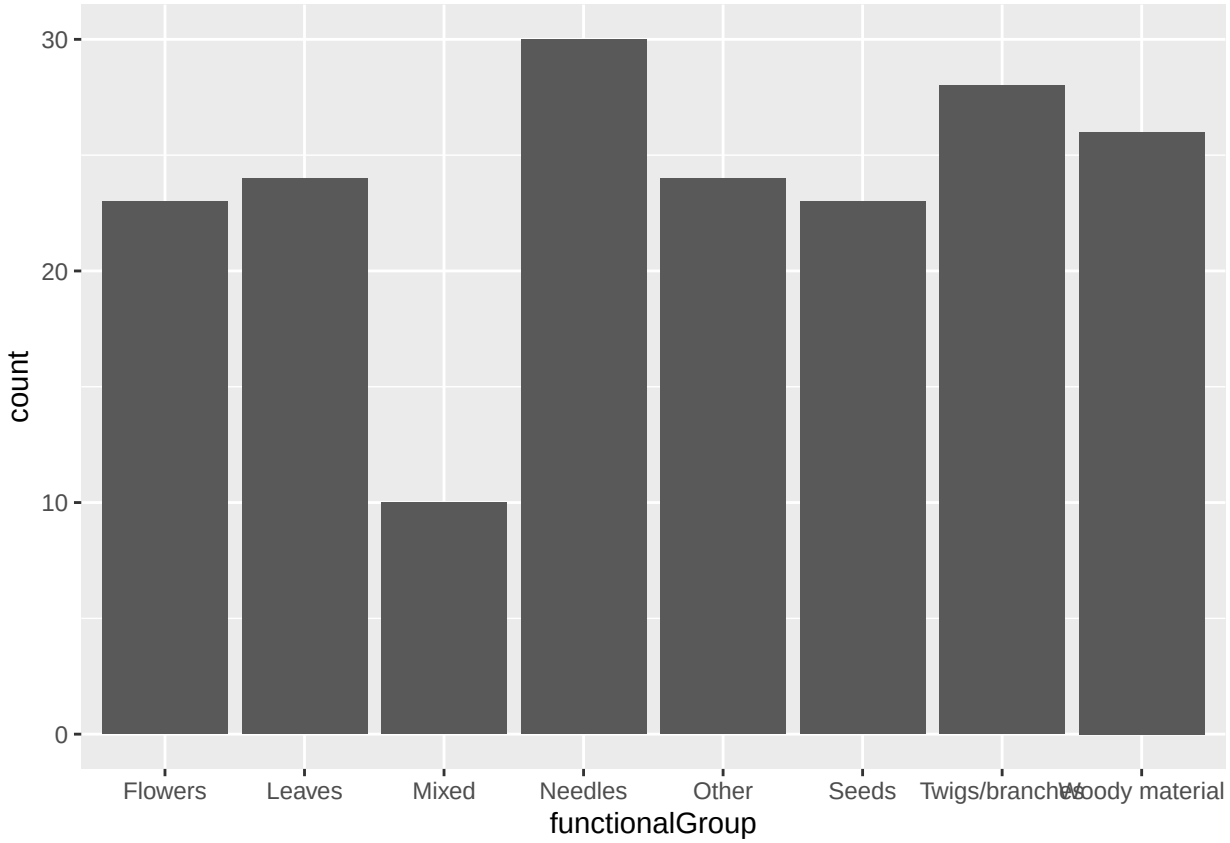
```
#view(litter.data$plotID)
```

Answer:

The unique function provides the exact unique number of different plots in the output as 12 levels, and the summary shows the different plots and how often they occurred and you would have to manually count to see how many different plots are there

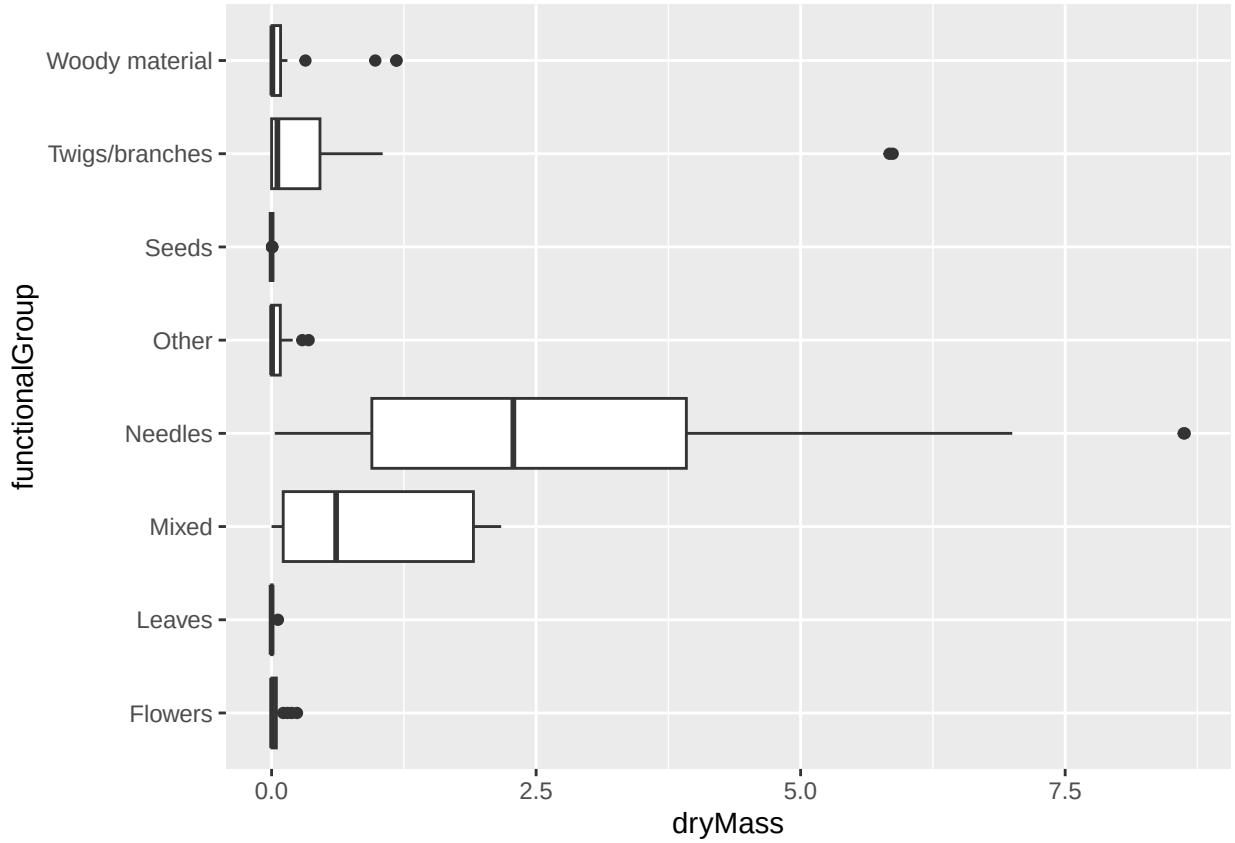
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(data = litter.data, aes(x = functionalGroup)) +
  geom_bar()
```

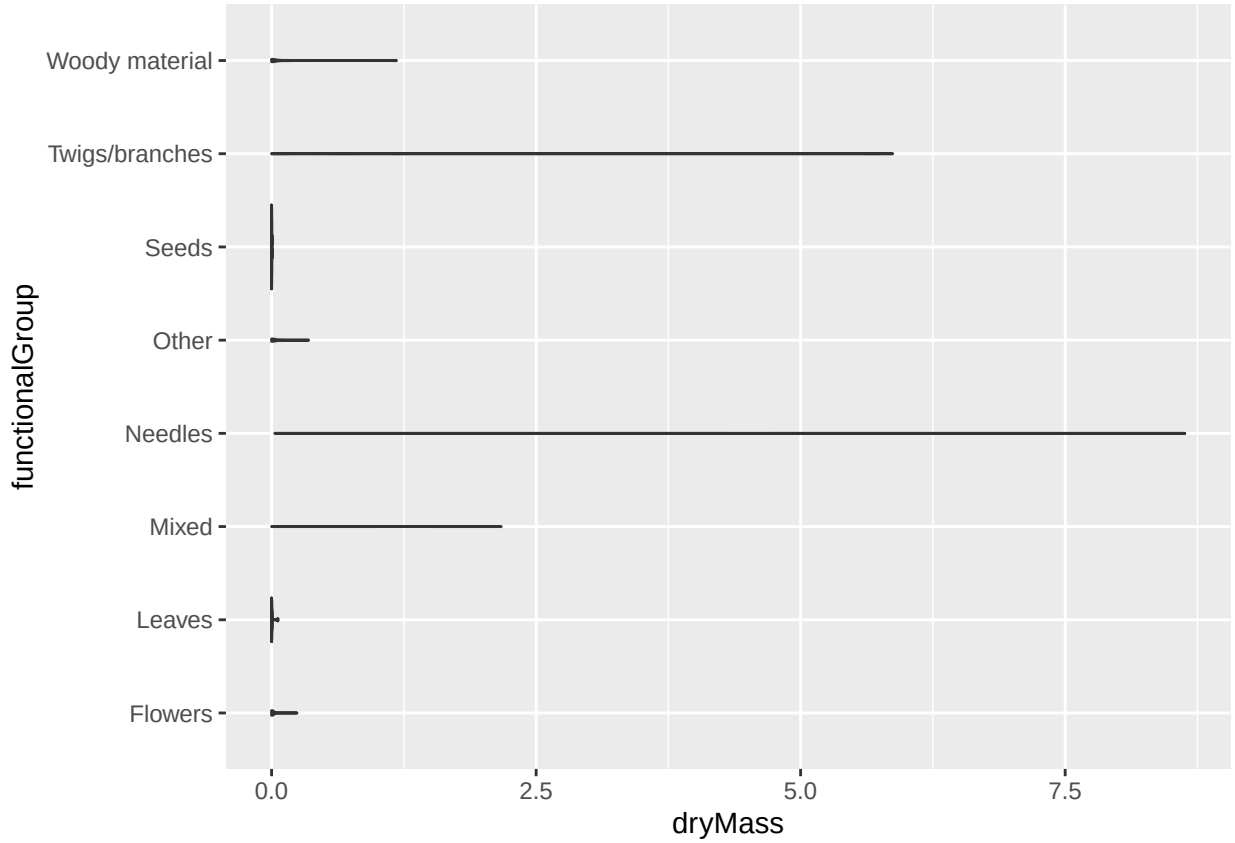


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(litter.data) +
  geom_boxplot(aes(x = dryMass, y = functionalGroup))
```




```
ggplot(litter.data) +
  geom_violin(aes(x = dryMass, y = functionalGroup),
              draw_quantiles = c(0.25, 0.5, 0.75))
```



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

Answer:

The boxplot is a more effectuve visualization option than violin becuse shows key summary statistics such as the median, quartiles (Q1 & Q3), range, and outliers, making it easier to interpret variability in dryMass across different functional groups.The violin plot is ineffective here due to skewed data and small sample sizes, leading to misleading density shapes.

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

Answer:

Needles and Mixed litter types have the highest biomass, with Needles showing the widest range and extreme outliers.