Weilin_Wang_Assignment3_DataExploration

Weilin Wang

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.

```
library(tidyverse)
library(lubridate)
library(here)

getwd()
```

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

```r
Neonics <- read_csv(here("Data", "Raw", "ECOTOX_Neonicotinoids_Insects_raw.csv"), col_types = cols(.defa

Litter <- read_csv(here("Data", "Raw","NEON_NIWO_Litter_massdata_2018-08_raw.csv"), col_types = cols(.d

head(Neonics)
```

```
## # A tibble: 6 x 30
##   ‘CAS Number‘ ‘Chemical Name‘           ‘Chemical Grade‘ Chemical Analysis Me~1
##   <fct>        <fct>                     <fct>            <fct>
## 1 58842209     Tetrahydro-2-(nitromethy~ Technical grade~ Unmeasured
## 2 58842209     Tetrahydro-2-(nitromethy~ Technical grade~ Unmeasured
## 3 58842209     Tetrahydro-2-(nitromethy~ Technical grade~ Unmeasured
## 4 58842209     Tetrahydro-2-(nitromethy~ Technical grade~ Unmeasured
## 5 58842209     Tetrahydro-2-(nitromethy~ Technical grade~ Unmeasured
## 6 58842209     Tetrahydro-2-(nitromethy~ Technical grade~ Unmeasured
## # i abbreviated name: 1: ‘Chemical Analysis Method‘
## # i 26 more variables: ‘Chemical Purity‘ <fct>,
## #   ‘Species Scientific Name‘ <fct>, ‘Species Common Name‘ <fct>,
## #   ‘Species Group‘ <fct>, ‘Organism Lifestage‘ <fct>, ‘Organism Age‘ <fct>,
## #   ‘Organism Age Units‘ <fct>, ‘Exposure Type‘ <fct>, ‘Media Type‘ <fct>,
## #   ‘Test Location‘ <fct>, ‘Number of Doses‘ <fct>,
## #   ‘Conc 1 Type (Author)‘ <fct>, ‘Conc 1 (Author)‘ <fct>, ...
```

```r
head(Litter)
```

```
## # A tibble: 6 x 19
##   uid               namedLocation domainID siteID plotID trapID weighDate setDate
##   <fct>             <fct>         <fct>    <fct>  <fct>  <fct>  <fct>     <fct>
## 1 7f065fec-bcb2-4~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-~ 2018-0~
## 2 88df210b-1445-4~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-~ 2018-0~
## 3 7f3c549c-1dfa-4~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-~ 2018-0~
## 4 97806ab5-42d2-4~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-~ 2018-0~
## 5 9d7c89f5-85f8-4~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-~ 2018-0~
## 6 6ca7a3e8-4d9e-4~ NIWO_061.bas~ D13      NIWO   NIWO_~ NIWO_~ 2018-08-~ 2018-0~
## # i 11 more variables: collectDate <fct>, ovenStartDate <fct>,
## #   ovenEndDate <fct>, fieldSampleID <fct>, massSampleID <fct>,
## #   samplingProtocolVersion <fct>, functionalGroup <fct>, dryMass <fct>,
## #   qaDryMass <fct>, remarks <fct>, measuredBy <fct>
```

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: Neonicotinoids are a class of systemic insecticides that affect the nervous systems of insects, leading to paralysis and death. Their use has been linked to Pollinator Decline. Pollinators like bees play a critical role in agriculture by pollinating crops, and their decline can have significant consequences for food production and biodiversity. The second effect could be water and Soil Contamination.These chemicals can persist in the environment, contaminating water bodies and soil, thus affecting aquatic insects and other invertebrates as well.

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 in forests is crucial for understanding nutrient cycling, carbon storage, and ecosystem dynamics. As organic matter like leaves, twigs, and fallen branches decompose, they release nutrients back into the soil, which supports plant growth and maintains soil fertility. Additionally, woody debris plays a key role in carbon sequestration, acting as a temporary carbon sink and influencing the forest's capacity to mitigate climate change, while also providing habitat for various organisms and contributing to overall 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.Litter Traps 2.Sampling Frequency 3. Woody Debris Transects

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

```
dim(Litter)
```

```
## [1] 188  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…]

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

sorted_effect_summary
```

```
##       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 
##     Intoxication     Accumulation     Biochemistry          Cell(s) 
##               12               12               11                9 
##       Physiology        Histology       Hormone(s) 
##                7                5                1
```

Answer:These effects are of particular interest because they directly influence insect population dynamics, survival rates, and their ecological roles. Studying mortality helps assess the immediate lethality of neonicotinoids, while behavioral changes and reproductive impairment provide insights into sub-lethal, long-term impacts that may affect insect populations and ecosystem balance over time.

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(Neonics$`Species Common Name`, maxsum = 6)
```

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

Answer:The six most commonly studied species might include honey bees, houseflies, mosquitoes, fruit flies, aphids, and beetles. These species are of interest because they either serve crucial ecological roles or are major agricultural pests.

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

```
## Warning: Unknown or uninitialised column: ‘Conc 1 (Author).‘.

## [1] "NULL"
```

```r
head(Neonics$`Conc 1 (Author)`)
```

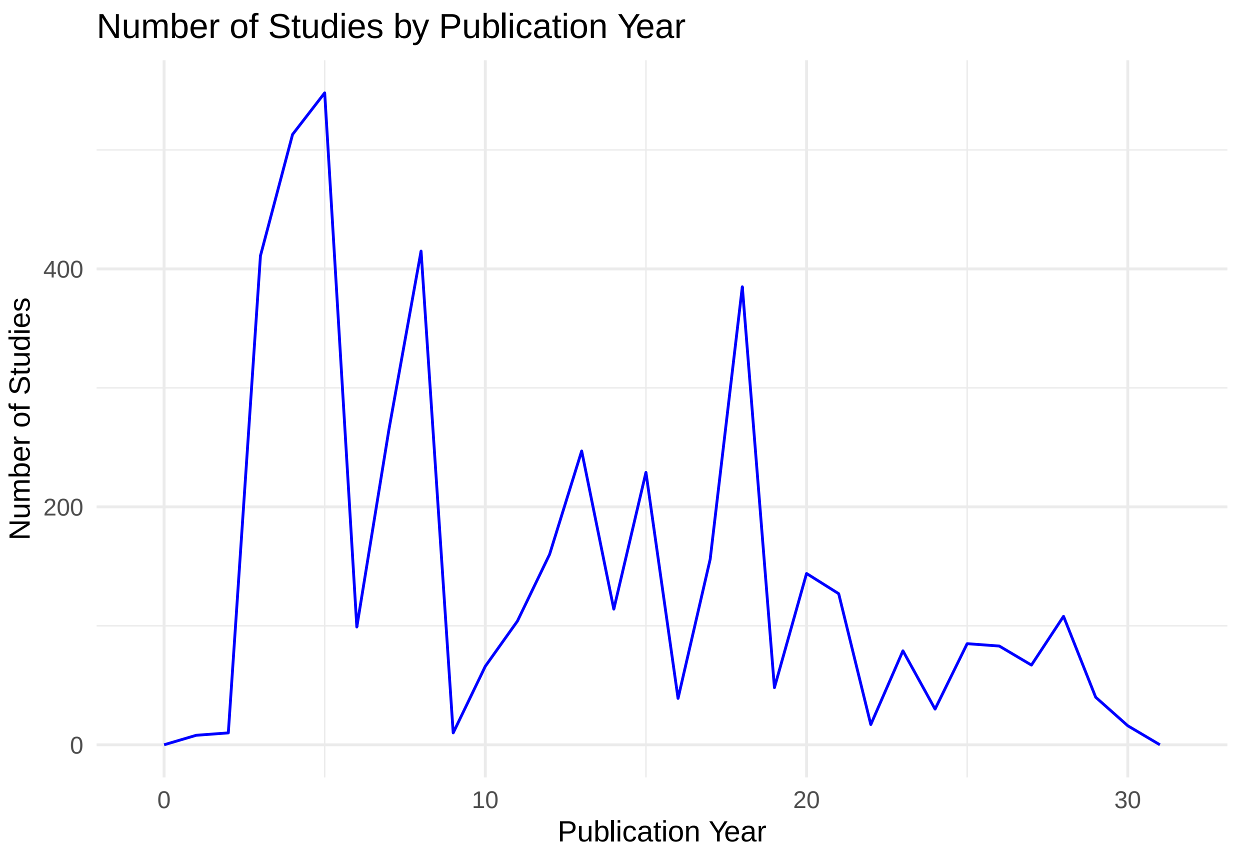
```
## [1] 27.2 19.7 47   25   13   268
## 1006 Levels: 27.2 19.7 47 25 13 268 170 28 48 40 83 900 15.3 20.4 5 NR ... 3.4859
```

Answer:column may not be numeric because it could contain non-numeric characters such as units or symbols like "<" indicating less than or greater than concentrations. These characters would prevent R from recognizing the column as numeric.

Explore your data graphically (Neonics)

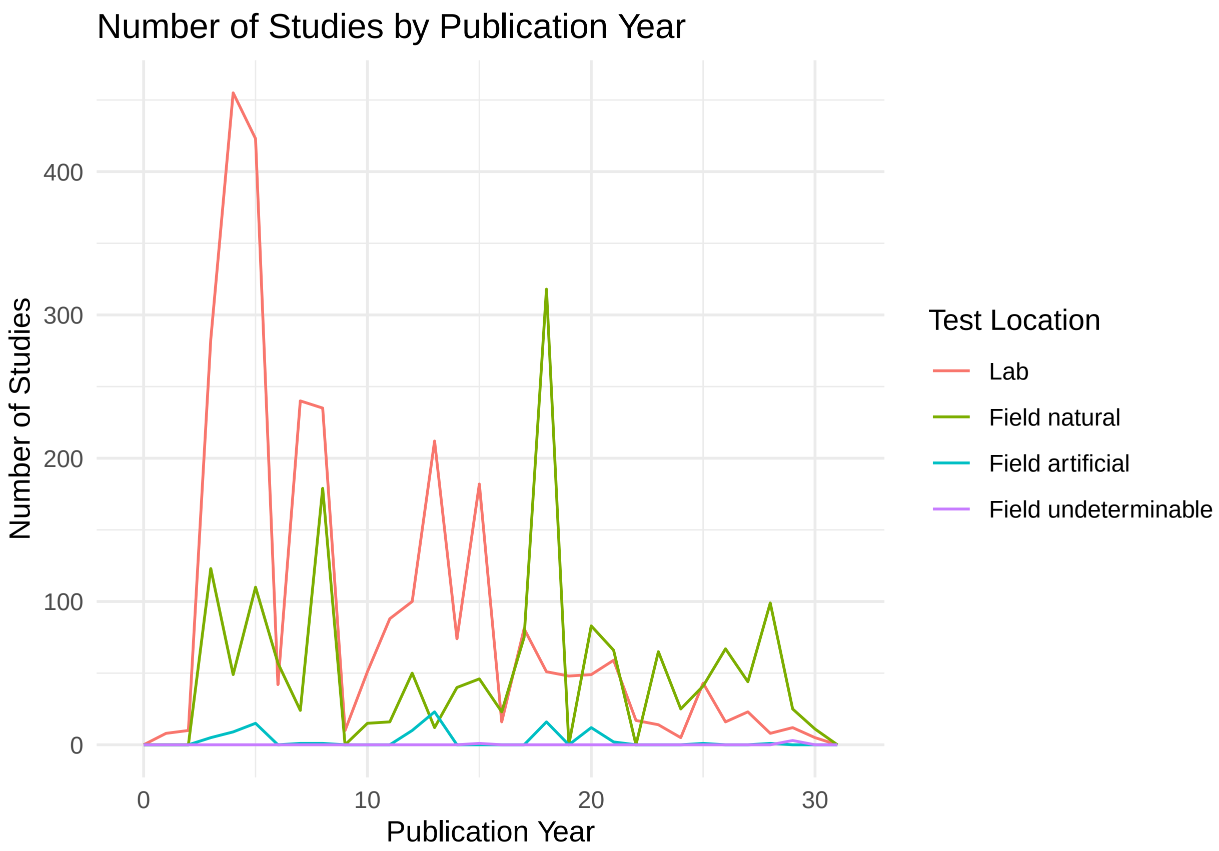
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```r
library(ggplot2)

ggplot(Neonics, aes(x = as.numeric(`Publication Year`))) +
  geom_freqpoly(binwidth = 1, color = "blue") +
  labs(title = "Number of Studies by Publication Year",
       x = "Publication Year",
       y = "Number of Studies") +
  theme_minimal()
```



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

```
ggplot(Neonics, aes(x = as.numeric(`Publication Year`), color = `Test Location`)) +
  geom_freqpoly(binwidth = 1) +
  labs(title = "Number of Studies by Publication Year",
       x = "Publication Year",
       y = "Number of Studies") +
  theme_minimal()
```



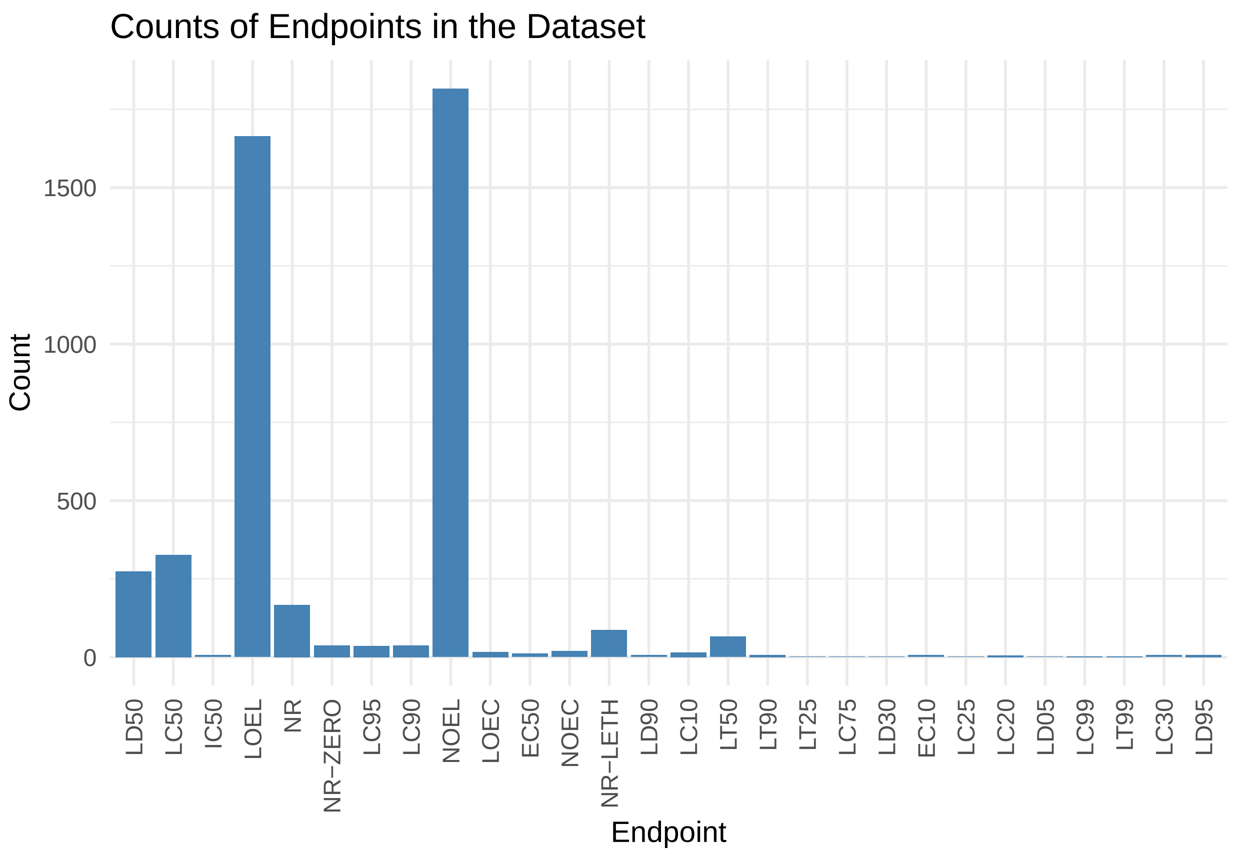
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer:The graph shows that laboratory studies are the most common, especially in the earlier years, while natural field studies become more prevalent over time. There is a noticeable shift from lab-based research toward more field-based studies, suggesting a growing interest in real-world ecological impacts. Studies conducted in artificial or undeterminable field environments remain consistently low throughout the observed period.

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, aes(x = `Endpoint`)) +
  geom_bar(fill = "steelblue") +
  labs(title = "Counts of Endpoints in the Dataset",
       x = "Endpoint",
       y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer:because they help determine the levels at which substances start causing observable harm (LOEL) and provide a measure of the substance's toxicity (LC50), which is essential for understanding the risk posed by neonicotinoids to insect populations.

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)

class(Litter$collectDate)
```

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

```
unique(Litter$collectDate[format(Litter$collectDate, "%Y-%m") == "2018-08"])
```

```
## [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`?

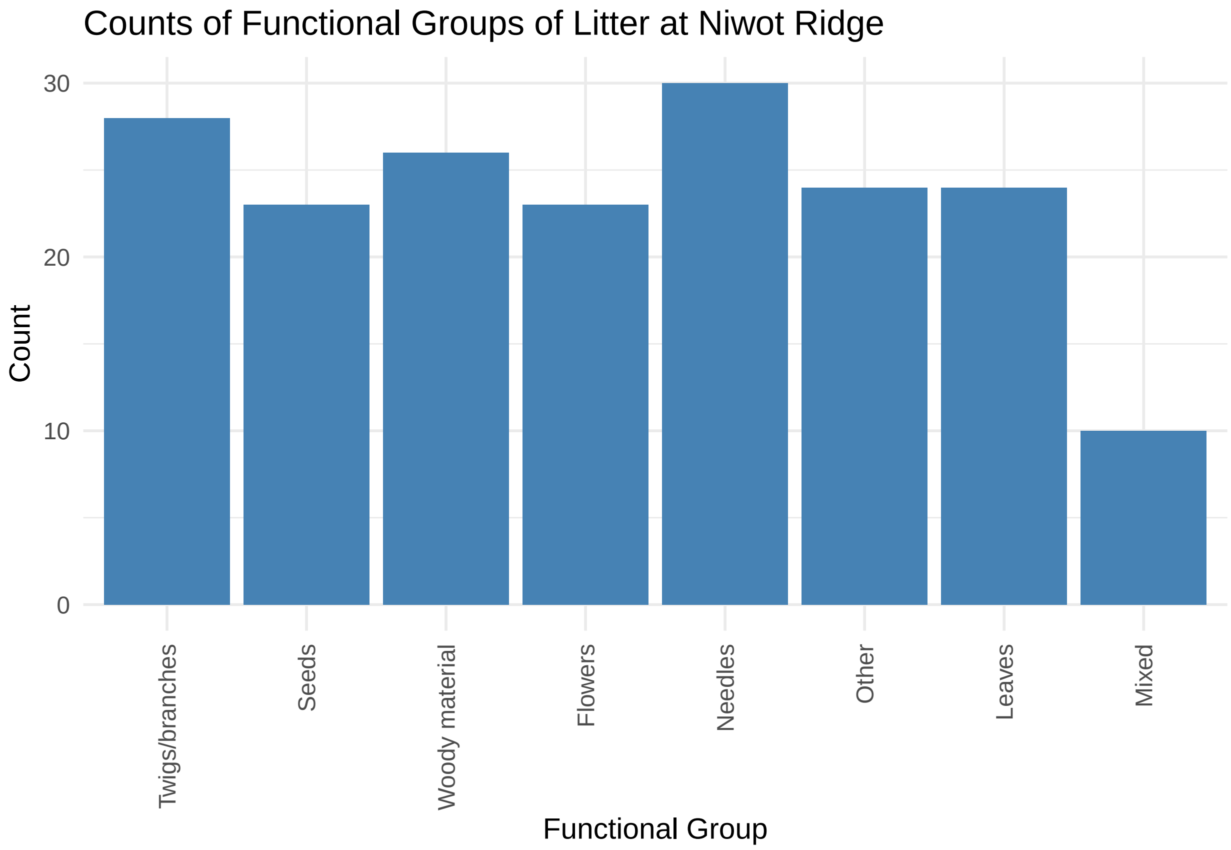
```
unique_plots <- unique(Litter$plotID)

length(unique_plots)
```

```
## [1] 12
```

Answer:unique helps you directly identify how many distinct plots were sampled, while summary would give the counts of how frequently each plot appears in the dataset but not directly the number of unique plots

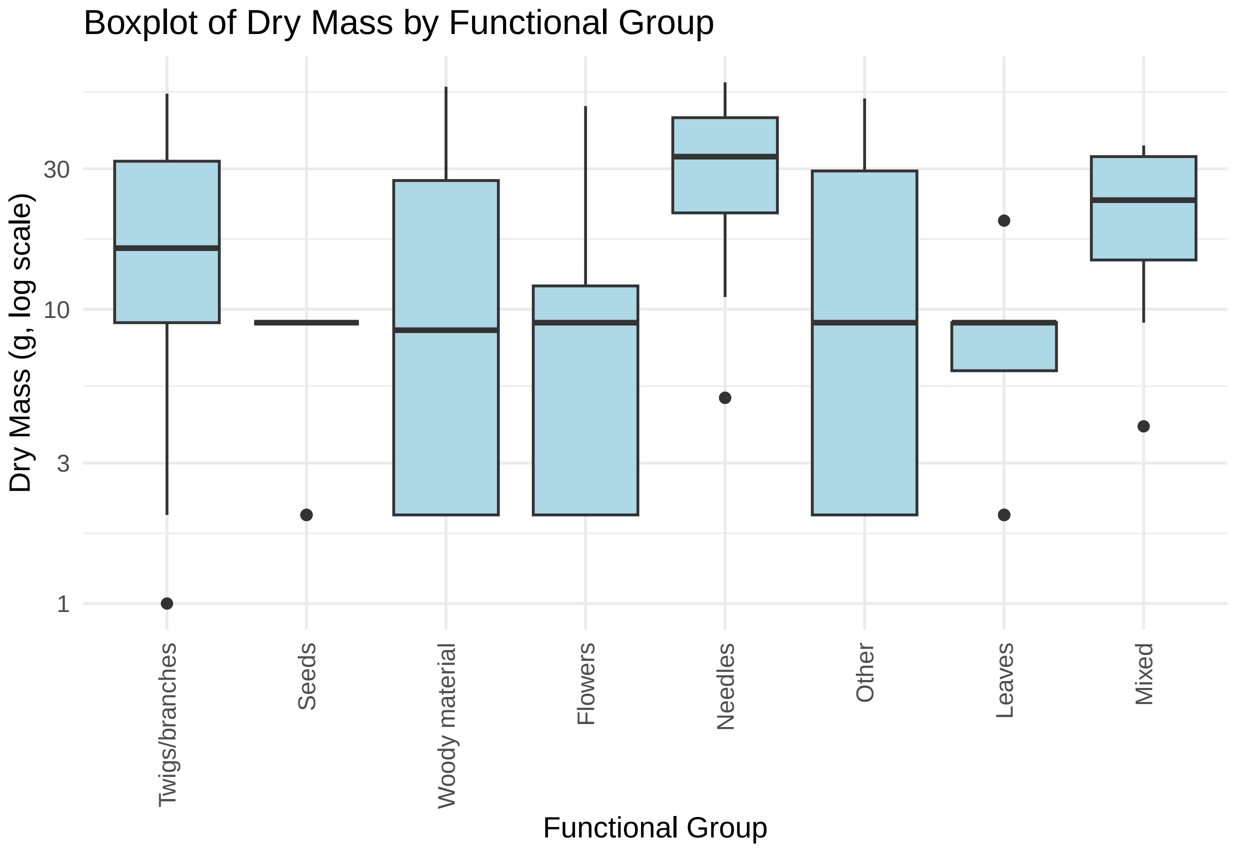
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar(fill = "steelblue") +
  labs(title = "Counts of Functional Groups of Litter at Niwot Ridge",
       x = "Functional Group",
       y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

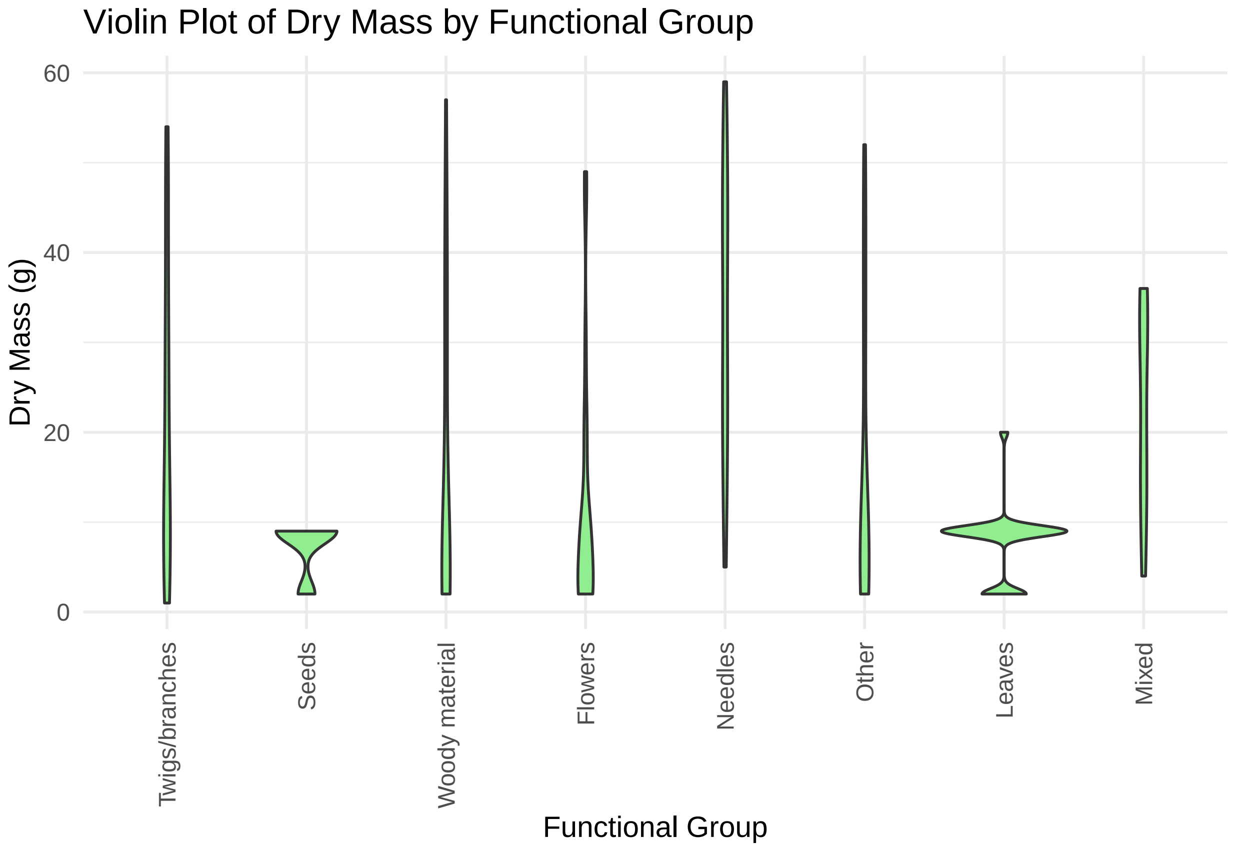


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
Litter$dryMass <- as.numeric(Litter$dryMass)

ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_boxplot(fill = "lightblue") +
  scale_y_log10() +  # Use log10 scale for better visualization if the range of dryMass is large
  labs(title = "Boxplot of Dry Mass by Functional Group",
       x = "Functional Group",
       y = "Dry Mass (g, log scale)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



```
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
  geom_violin(fill = "lightgreen") +
  labs(title = "Violin Plot of Dry Mass by Functional Group",
       x = "Functional Group",
       y = "Dry Mass (g)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



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 median, quartiles, and potential outliers of the dryMass distribution for each functional group, which are essential for comparing central tendencies and data spread. The violin plot, while useful for showing the distribution's shape, can be harder to interpret when comparing specific statistical summaries across groups.

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

> Answer:twigs/branches and woody material tend to have the highest biomass at these sites. The upper ranges for these functional groups show higher dry mass values compared to other types of litter like seeds, flowers, and needles. This suggests that larger, denser materials such as branches and woody debris contribute more significantly to litter biomass at Niwot Ridge.