Assignment 3: Data Exploration

Merve Onal

Spring 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 Sakai.

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. Check your working directory, load necessary packages (tidyverse, lubridate), 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.

```
#First step is to check my working directory
getwd()
```

```
## [1] "/home/guest/EDA_Spring2024"
```

```
#I load the packages
#Here I import the datasets,
#I name these datasets according to the directions of the assignment
#I include the subcommand to read strings in as factors
#I install Packages
library(tidyverse)
library(lubridate)
Neonics <- read.csv("~/EDA_Spring2024/Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",stringsAsFactors =
Litter <- read.csv("~/EDA_Spring2024/Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv",stringsAsFacto
```

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:Based on information available online, neonicotinoid have been widely used in agriculture to protect crops from various pests. Neonicotinoids exhibit high acute toxicity to a broad range of insects, including both target and non-target species. Bees and other pollinators have been the focus of many studies due to their crucial role in pollination. Besides, these insects which are killed by these pesticides are food sources for other species.

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:Trees that topple and decompose within the forest contribute nutrients to the soil and help retain moisture. Wood that has fallen and is larger than 7 cm in diameter is known as coarse woody debris. The decomposition rate of this woody litter is influenced by factors such as moisture and temperature. This litter and debris serve as both sustenance and shelter for a diverse array of organisms, ultimately enhancing biodiversity. The quantity of coarse woody debris is a crucial factor in assessing and restoring temperate deciduous forests.

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 0.5 m square traps and PVC baskets elevated ~80cm above the ground. Fine woody debris is collected from 3 m x 0.5 m rectangular ground traps. 2.In sites with forested tower airsheds, trap placement is randomized from a grid of possible locations. In non-forested sites, trap placement is targeted beneath areas of woody vegetation. 3. Traps are sampled at varying frequencies depending on the site. Deciduous forests are sampled more frequently during leaf fall. Each trap collects litter sorted into functional groups like leaves,needles, twigs, woody materials, seeds, etc.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

Here we see the dimensions of the dataset

```
dim(Neonics)
```

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

```
dim
```

```
## function (x)  .Primitive("dim")
```

4623 indicates the number of rows in the dataset,whereas 30 shows the number of columns

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?

```
#I use the "summary" function on the effect column
#I print the summary to see the most common effects
effect_summary <- summary(Neonics$Effect)
print(effect_summary)
```

```
##     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: Population (1803 occurrences) is mostly studied to understand how these pesticides influence the abundance and dynamics of insect populations. The change in population could give harm to the ecosystem. Mortality (1493 occurrences) is the second most studied concern because of the direct and immediate impact of neonicotinoid pesticides on insect populations. High mortality rates can cause to population declines, potentially affecting crop pollination and ecosystem balance.Behavior (360 occurrences) is also studied because they can affect an insect's ability to survive and reproduce. Behavioral disruptions can lead to reduced efficiency in resource acquisition and survival challenges.

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: The `sort()` command can sort the output of the summary command...]

```
#I use the summary function to determine the column "common name"
#I sort the summary results in descending order to see the values/categories that appear most often
#Then I select the top six most commonly studied species in the dataset
#Finaly, I print the top species
species_counts <- summary(Neonics$Species.common.name)
sorted_species_counts <- sort(species_counts, decreasing = TRUE)
top_species <- head(sorted_species_counts, n = 6)
print(top_species)
```

```
##  Class   Mode Length
##   NULL   NULL      0
```

Answer:These species have several things in common. They are crucial pollinators, support plant reproduction, and food production, and they are found in diverse ecosystems. Consequently, these species are convenient for ecological research, and their presence in agricultural and urban settings makes them suitable for studies on human impacts on ecosystems. There is a large literature about these species which facilitates research. Especially Honey Bees play a vital role in crop pollination and contribute significantly to the economy. Declines in pollinator populations emphasize the need for conservation efforts.

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?

```
class("Conc.1..Author.")
```

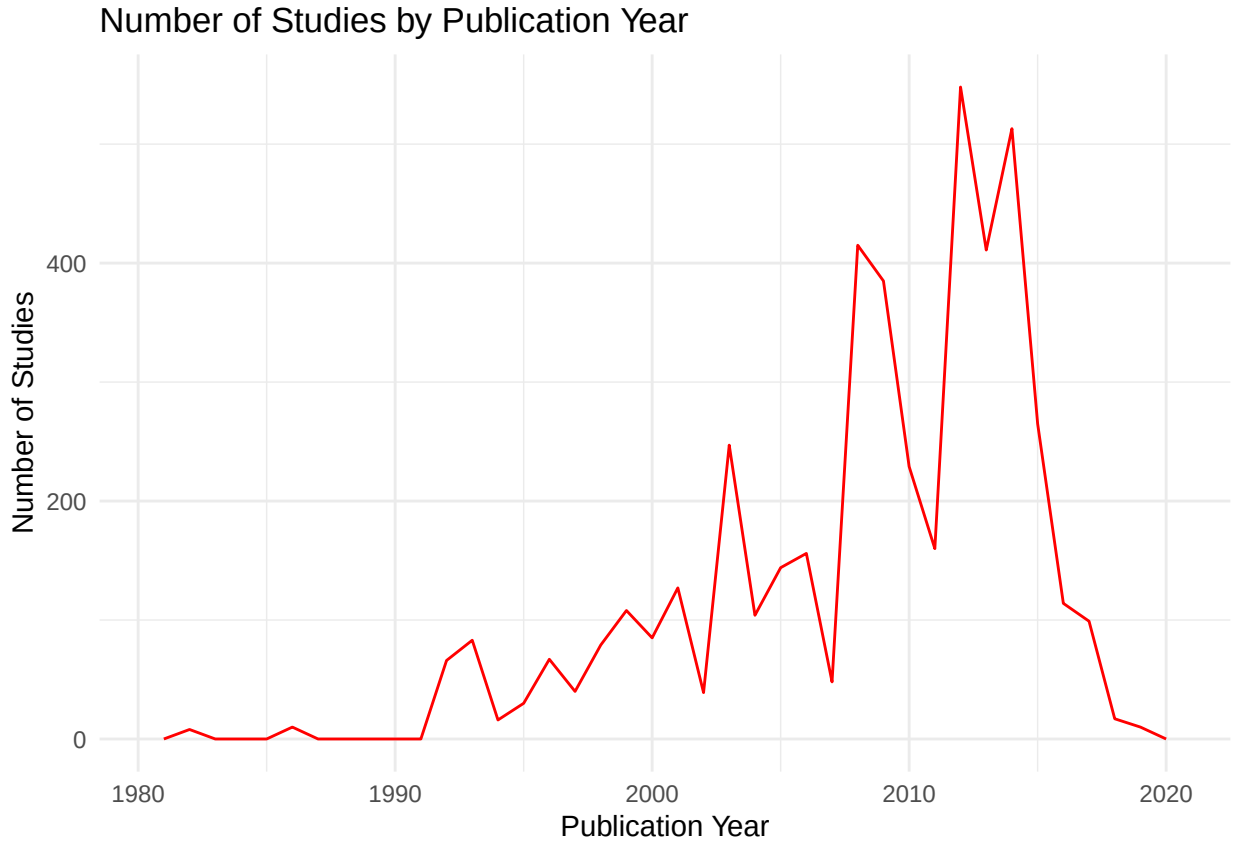
```
## [1] "character"
```

Answer:The class of "Conc.1..Author." is character. It is not numeric in our dataset because this column contains non-numeric characters or symbols.

Explore your data graphically (Neonics)

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

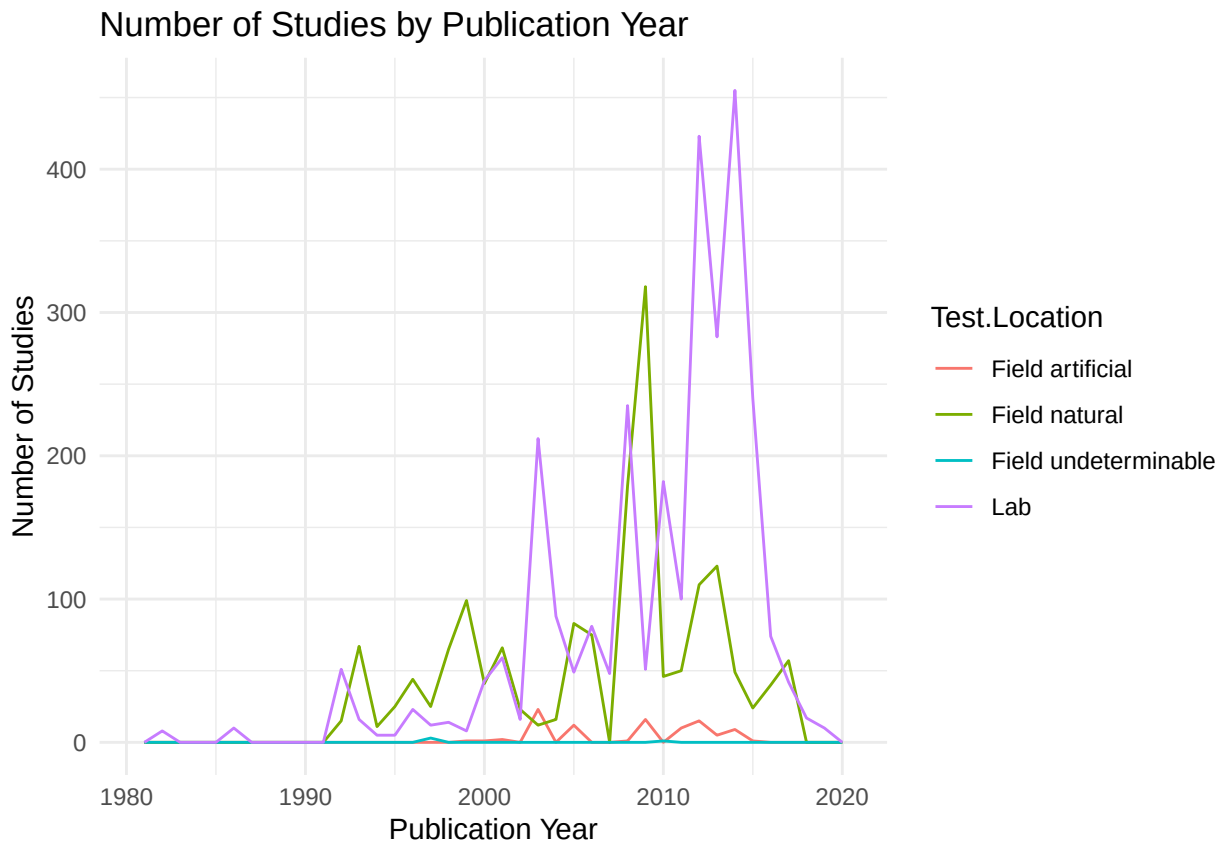
```
#I load the package from library
library(ggplot2)
#I create the plot of the number of studies
ggplot(data = Neonics, aes(x = Publication.Year)) +
geom_freqpoly(binwidth = 1, color = "red") +
labs(title = "Number of Studies by Publication Year",
x = "Publication Year",
y = "Number of Studies") +
theme_minimal()
```



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

```
ggplot(data = Neonics, aes(x = 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?

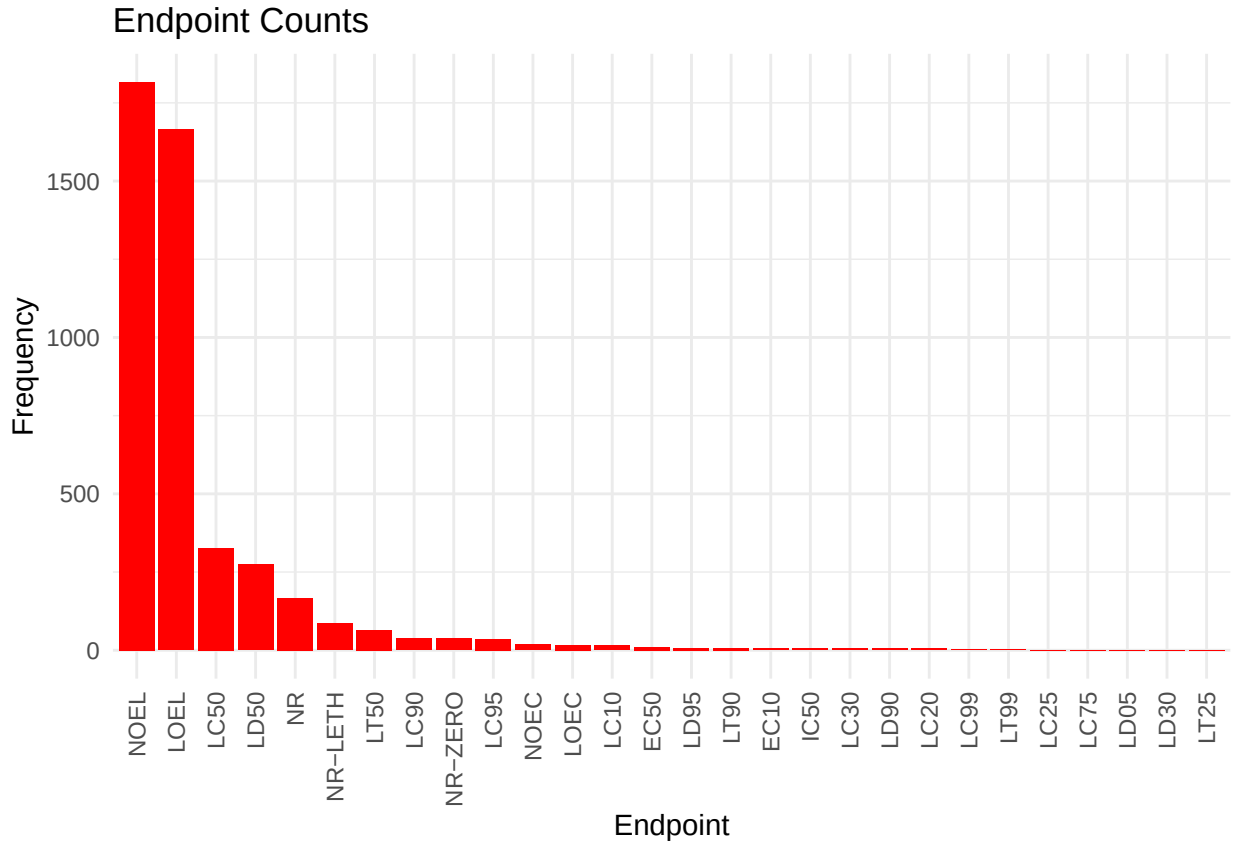
> Answer:The most common test locations are Lab, and as seen in the graph the number of studies conducted in Lab increase significantly by the time. In addition to this, the second most common test location is Field natural, which reaches its peak just before the year 2010 and becomes the most common test location of the year. Since 2010, the test conducted in the Field natural is in a declining trend.

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...]

```
library(dplyr)
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
## filter, lag
## The following objects are masked from 'package:base':
## intersect, setdiff, setequal, union
# I calculate the counts of each unique endpoint
endpoint_counts <- table(Neonics$Endpoint)
# We convert the counts to a data frame for plotting
endpoint_counts_df <- data.frame(
```

```
Endpoint = names(endpoint_counts),
Frequency = as.numeric(endpoint_counts))
# Here we sort the data frame by frequency in descending order
endpoint_counts_df <- endpoint_counts_df %>%
arrange(desc(Frequency))
# I create the bar graph
ggplot(data = endpoint_counts_df, aes(x = reorder(Endpoint, -Frequency), y = Frequency)) +
geom_bar(stat = "identity", fill = "red") +
labs(title = "Endpoint Counts",
x = "Endpoint",
y = "Frequency") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



Answer:As it is shown in the graph, the two most common endpoints are NOEL (No Observed Effect Level) and LOEL (Lowest Observed Effect Level). NOEL and LOEL are commonly used in risk assessments and research. NOEL is The highest exposure level at which there are no effects (adverse or nonadverse) observed in the exposed population, when compared with its appropriate control. LOEL is the is the lowest dosage level at which chronic exposure to the substance shows adverse effects. There is a large literatıre criticize these two tools of being flawed.

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.

```
# Here I check the class of the collectDate variable
class(Litter$collectDate)
```

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

```
## [1] "factor"
# Since it was a factor, now I change it to a date
Litter$collectDate <- as.Date(Litter$collectDate)
#Check the class of the variable after the change and it is now a date
class(Litter$collectDate)
```

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

```
## [1] "Date"
#Now I use the unique function to find dates when litter was sampled in August 2018
unique_dates <- unique(Litter$collectDate)
august_2018_dates <- unique_dates[format(unique_dates, "%Y-%m") == "2018-08"]
august_2018_dates
```

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

```
## [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)
```

```
## NULL
```

```
print(Litter$PlotID)
```

```
## NULL
```

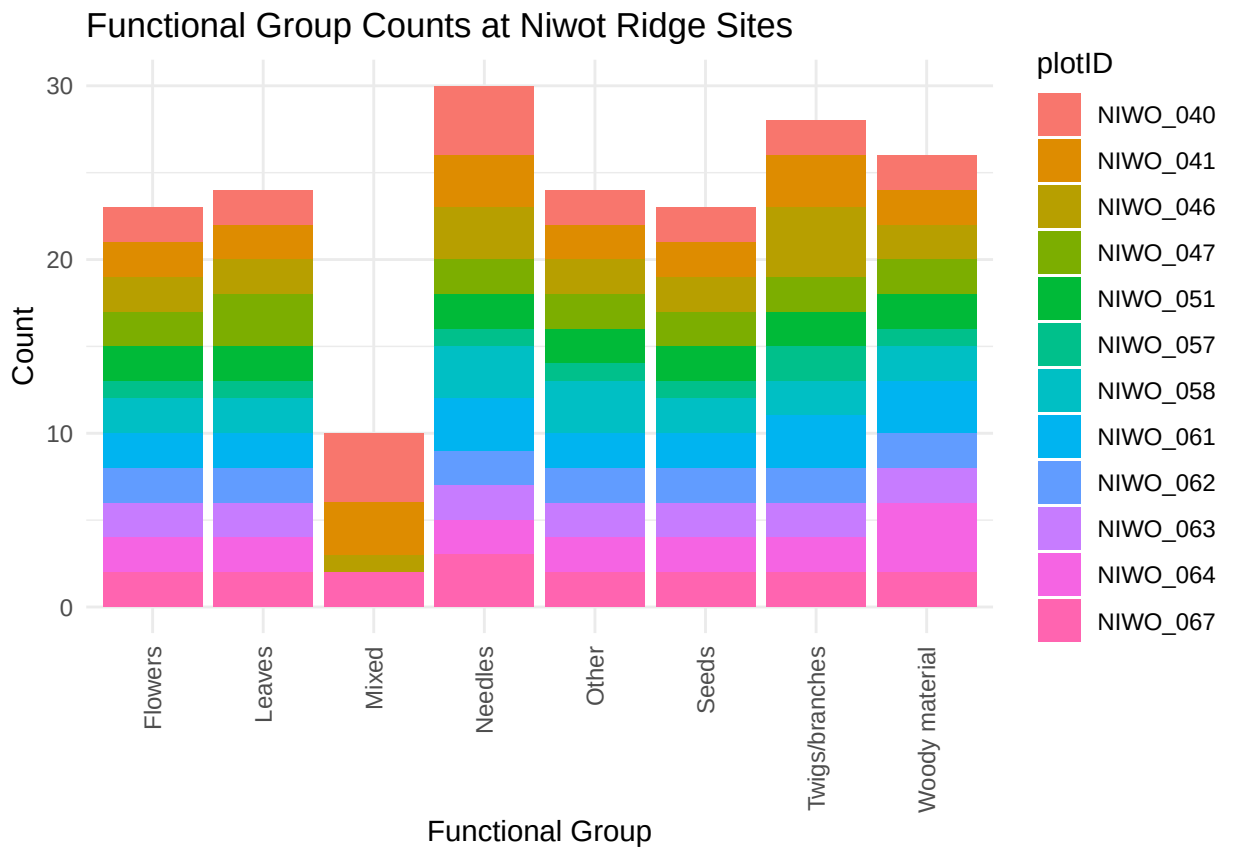
```
unique_plots <- unique(Litter$PlotID)
num_plots <- length(unique_plots)
num_plots
```

```
## [1] 0
```

Answer: The unique() function in R is used to eliminate or delete the duplicate values or the rows present in the vector, data frame or matrix. The unique function extracts the unique elements. The summary function provides a summary of the central tendency, dispersion, and distribution of a given set of values. The summary function is also could be used to exhibit the unique values.

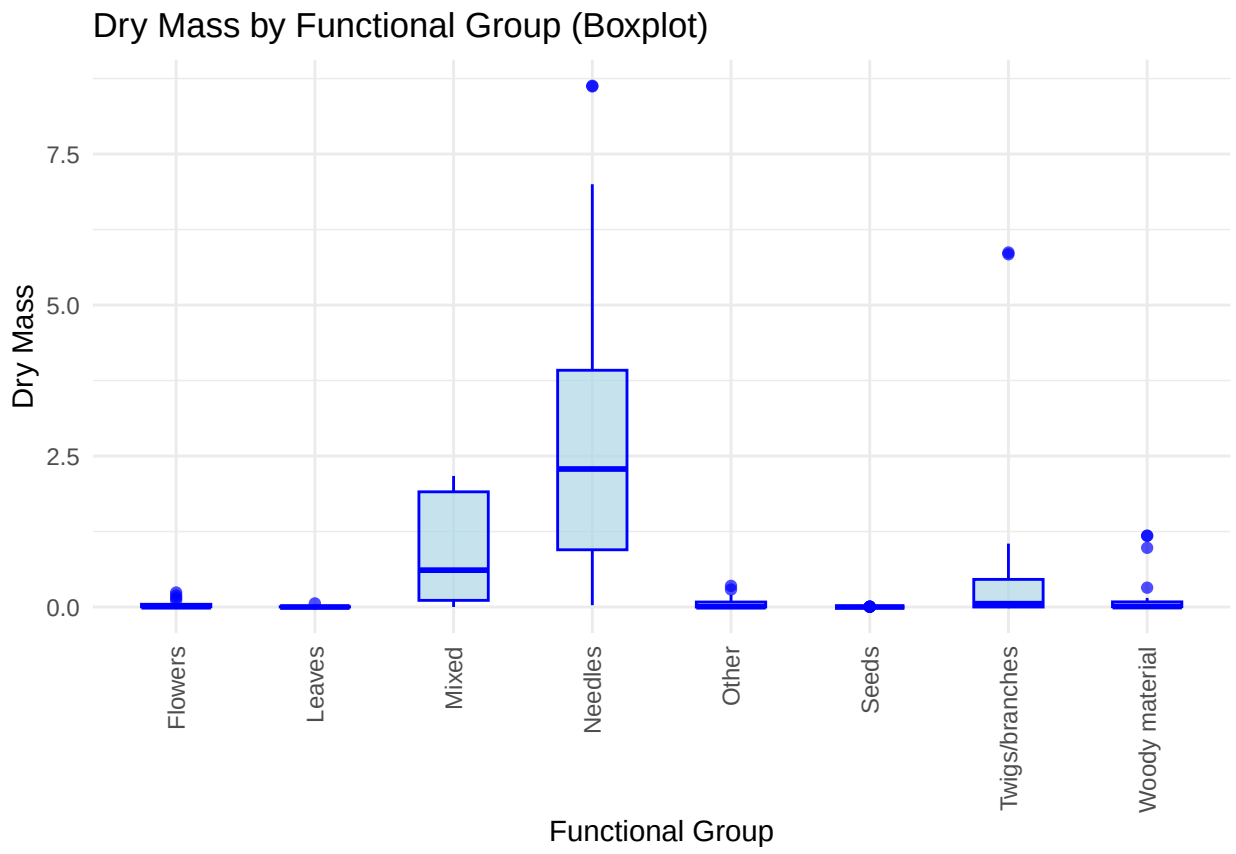
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

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

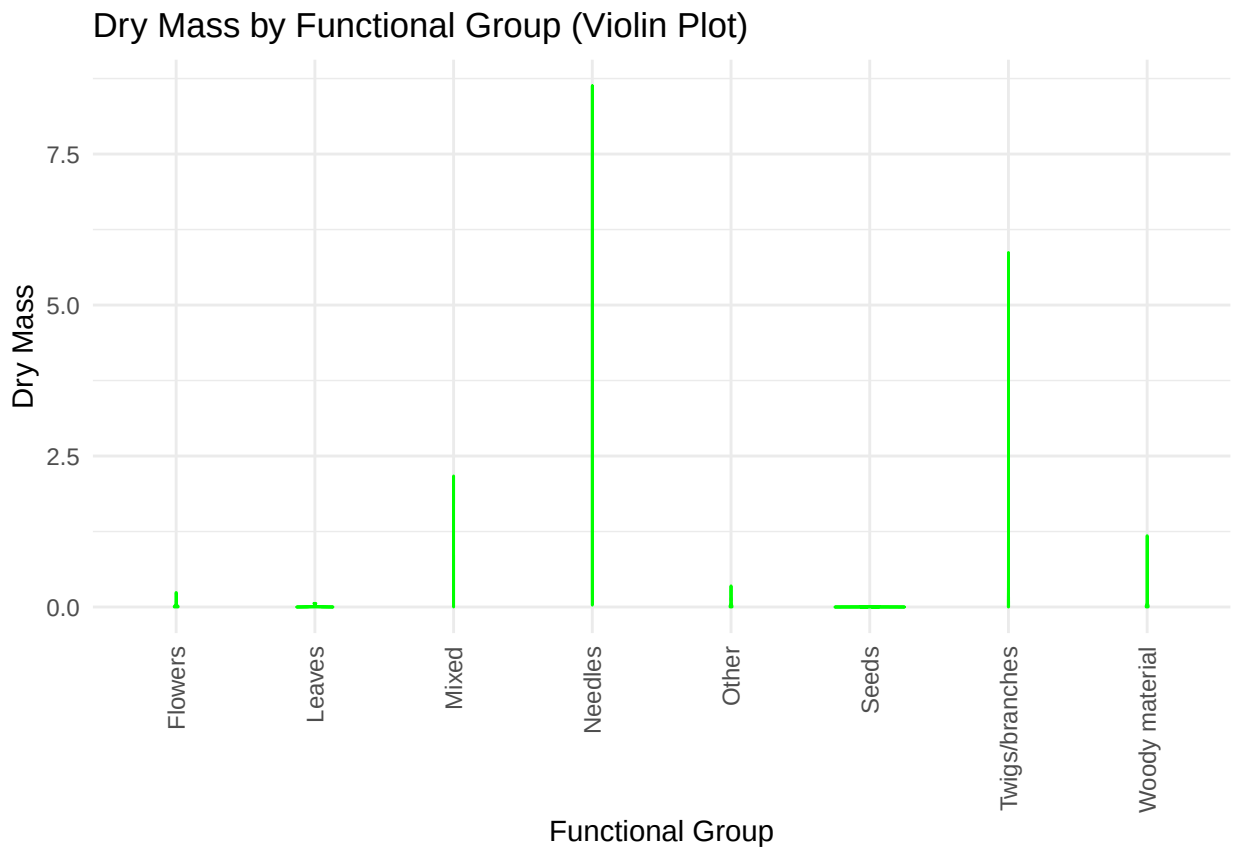


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
# Boxplot for dryMass by functionalGroup
boxplot_plot <- ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
geom_boxplot(fill = "lightblue", color = "blue", alpha = 0.7, width = 0.5) +
labs(title = "Dry Mass by Functional Group (Boxplot)",
x = "Functional Group",
y = "Dry Mass") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
print(boxplot_plot)
```




```r
# Violin plot for dryMass by functionalGroup
violin_plot <- ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
geom_violin(fill = "lightgreen", color = "green", alpha = 0.7, width = 0.5) +
labs(title = "Dry Mass by Functional Group (Violin Plot)",
x = "Functional Group",
y = "Dry Mass") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
print(violin_plot)
```



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

Answer:In this case, we examine the dispersion of dry mass values among different functional groups. The boxplot provides insights into the central tendency, with the horizontal line denoting the median for each specific functional group. Additionally, the height of each box illustrates the interquantile range, indicating the spread. Notably, outliers are observable in the plot, evident as individual data points falling beyond the expected range. Whereas,the violin plot contains limited information and does not indicate enough observations to make a comment. Consequently, we can say that boxplot has more effective visualization than the violin plot.If I have many points, then the box plot may not be very insightful,so it may require to try violin plot.

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

Answer: The groups with the highest median dry mass(horizontal line) tend to have the highest biomass on average. The Needles and Twigs/branches has the highest biomass.