Assignment 3: Data Exploration

Andi Mujollari

Fall 2023

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.

```
#Here i set up the working directory
getwd()
```

```
## [1] "/Users/andimujollari/Desktop/EDE-Fall2023"
```

```
#Here i installed the packages that will help me in working with data
#install.packages(tidyverse)
#install.packages
#library(tidyverse)
```

```
#install.packages(lubridate)
#install.packages
#library(lubridate)
```

```
Neonics.data <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = TRUE)
Litter.data <- read.csv("./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.

   Answer: Neonics are a group of man-made, neurotoxic insecticides that are applied to lawns, gardens, golf courses, and pets to control flea and tick infestations. Neonics, which were created in the middle of the 1990s, are currently the most widely used kind of insecticide in the US.They function by adhering tenaciously to insects' nerve cells, overstimulating and decimating them. Exposed insects frequently spasm and shake uncontrollably, then become paralyzed before dying. Neonics can impair essential abilities of insects, including their immune system, navigation, stamina, memory, and fertilization, even at nonlethal levels.

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: It is essential to research litter and woody debris in forests because they provide insights into the complex processes of forest ecosystems. These organic components that have fallen to the ground are important in the cycling of nutrients and supply vital ingredients for plant development. Additionally, they provide a variety of creatures with habitats, enhancing biodiversity. Additionally, the dynamics of climate change are impacted by carbon storage due to the disintegration of this debris. It affects fire danger, aiding in the control of wildfires, and acts as a gauge of overall forest health, directing conservation initiatives and sustainable forest management techniques. Additionally, it helps prevent soil erosion and preserve the quality of the water.

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 litter sampling is concentrated on 20 40m*40m plots in locations with wooded tower airsheeds. 2.The litter sample is concentrated on 4 40m*40m tower plots and 26 20m*20m plots in locations with low-saturated vegetation above the tower airsheeds. 3. For every 400m2 plot area, one pair of traps is set up, giving each plot between 1-4 trap pairs.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#Through the 'dim' command i am assessing the simensions of the Neonics dataset
dim(Neonics.data)
```

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

```
dim
```

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

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?

```
#Here i find th emost common effect that are used in our dataset.
Effect_summary <- summary(Neonics.data$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: In my opinion, it is on our interest to undersant the vital dynamics of Nenonics. We use them as fertilizers but at the same time they have huge negative impacts on the population of bee's and other un-harmful insects. These most used effects are informing us regarding the vital dynamics of these species.

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...]

```
#Now i use the summary funciton to identify the common name
Neonics_summary <- summary(Neonics.data$Species.Common.Name)

#Here i sort my database in a decreasing order so i the machine can focus in the most common ones
sorted_Neonics_summary <- sort(Neonics_summary, decreasing = TRUE)
#Finally we select only the most common 6 of them
Common_species <- head(sorted_Neonics_summary, n=6)

print(Common_species)
```

```
##               (Other)             Honey Bee        Parasitic Wasp
##                   670                   667                   285
## Buff Tailed Bumblebee   Carniolan Honey Bee            Bumble Bee
##                   183                   152                   140
```

Answer: The majority of the six most common studied species in this dataset are part of the Bee family. According to bee experts at the Food and Agriculture Organization (FAO) of the United Nations, a third of the world's food production depends on bees. If this perticide is harming their population, future food security has to be considered.

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?

```
#Here i found the class of this column.
class('Conc.1..Author')
```

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

Answer: The class of the Conc.1..Author is character and not numerical because it contains non-numerical symbols for example 'NR'

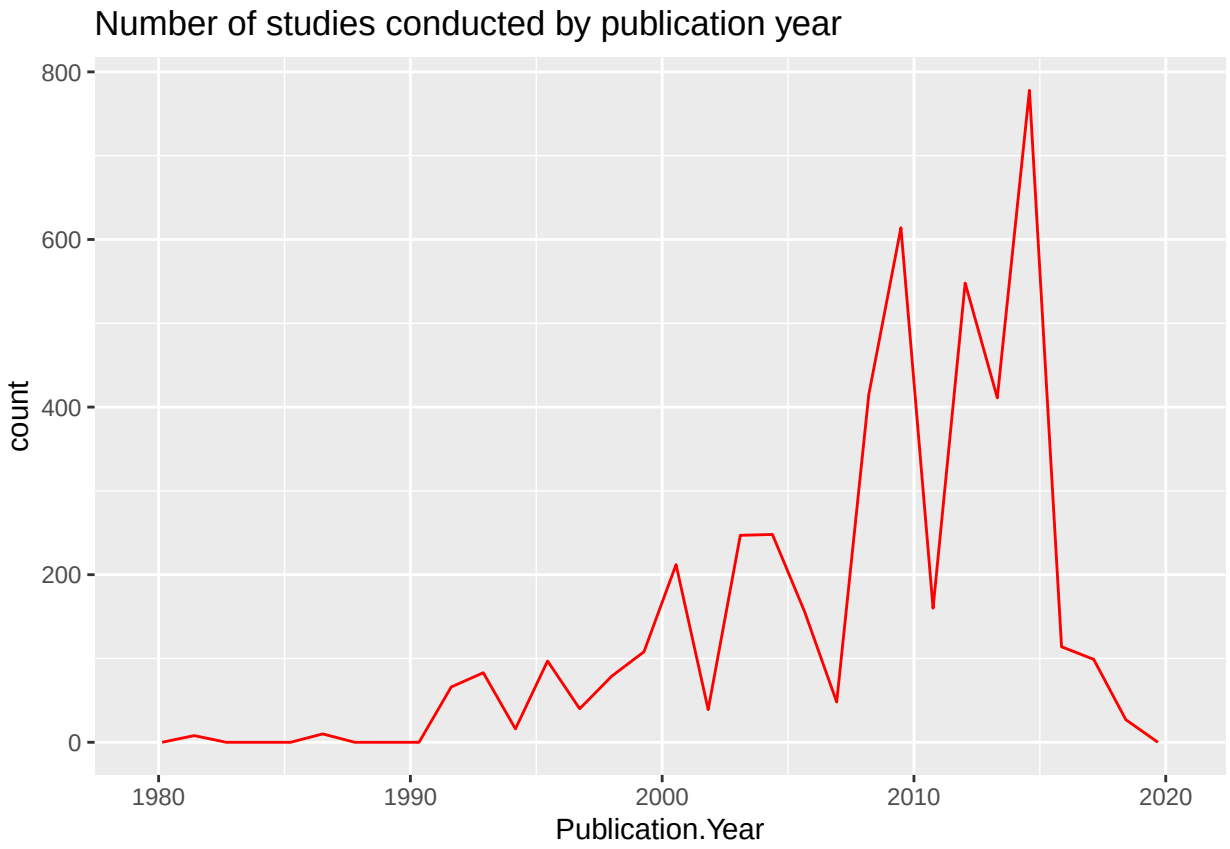
Explore your data graphically (Neonics)

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

```
#To create a visual presentation of my data, first i have to get from the ggplot package from library
library(ggplot2)
#Here i say to R what data i want to be presented, in which axes, the title of the graph, color...etc.
ggplot(data = Neonics.data, mapping = aes(x=Publication.Year)) +
  geom_freqpoly(bindwith = 1.5, color = "red") + labs(title = "Number of studies conducted by publicati
```

```
## Warning in geom_freqpoly(bindwith = 1.5, color = "red"): Ignoring unknown
## parameters: `bindwith`
```

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

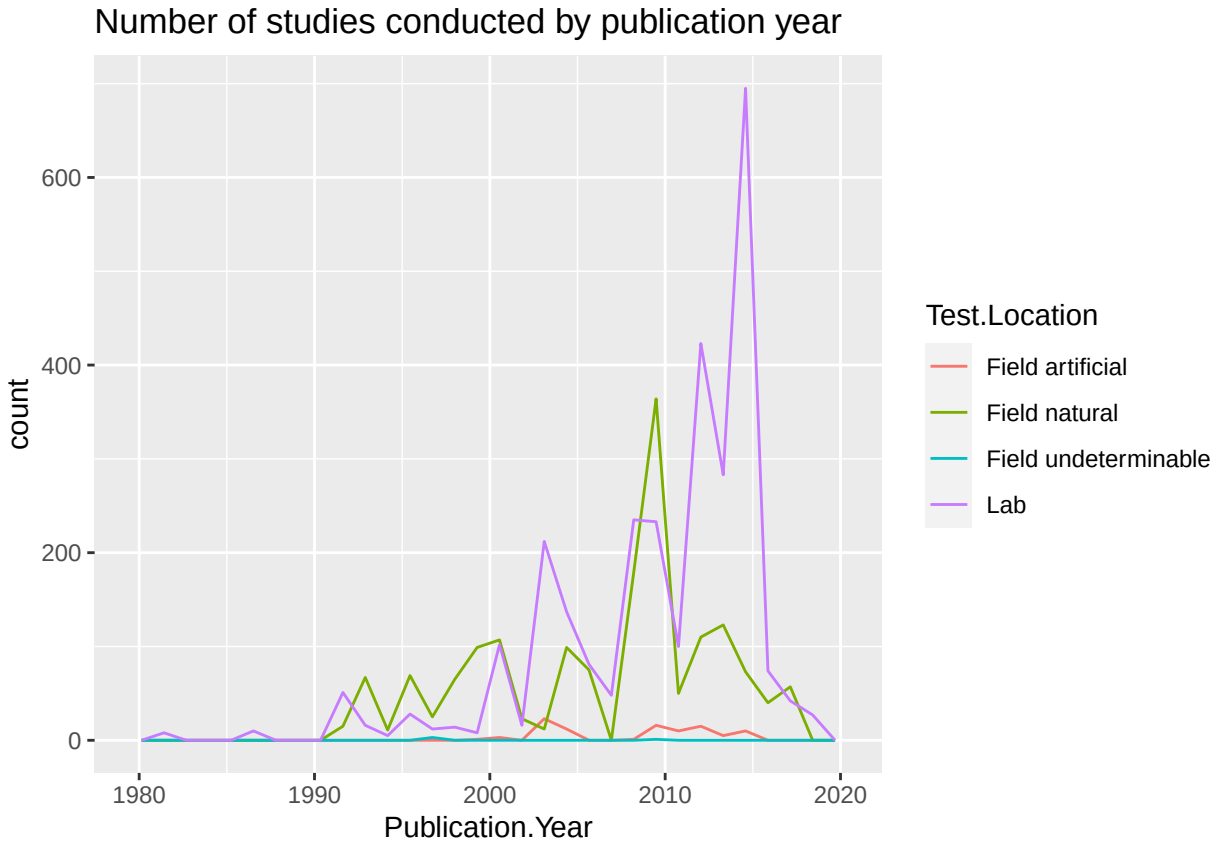


10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
#Now i will use ask R to create a graph with different colors representing different locations.
ggplot(data = Neonics.data, mapping = aes(x=Publication.Year, color = Test.Location)) +
  geom_freqpoly(bindwith = 1.5) + labs(title = "Number of studies conducted by publication year", X = "
```

```
## Warning in geom_freqpoly(bindwith = 1.5): Ignoring unknown parameters:
## `bindwith`
```

```
## `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: Acroding to the graph, it is shown that the most common test location especially during the last years are conducted in the labs. Yes, they are constantly changing over time.

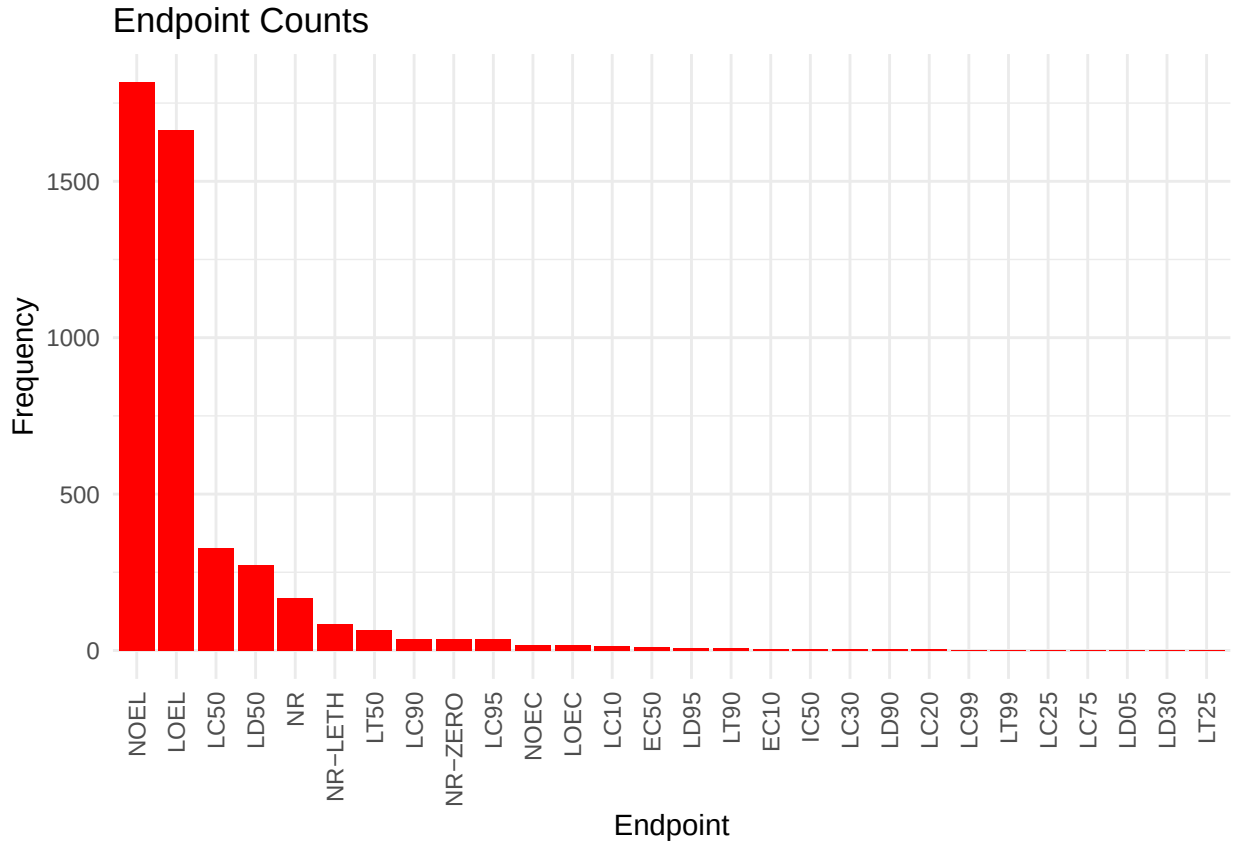
11. Create a bar graph of Endpoint counts. What are the two most common end points, and how are they defined? Consult the ECOTOX_CodeAppendix for more information.

[**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...]

```
#Here i have to get the dplyr package first

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
```

```r
# Here I calculate the counts of each unique endpoint
endpoint_counts <- table(Neonics.data$Endpoint)

# Here i convert the counts to a data frame for plotting
endpoint_counts_df <- data.frame(
  Endpoint = names(endpoint_counts),
  Frequency = as.numeric(endpoint_counts))

# Now i sort the data frame by frequency in descending order
endpoint_counts_df <- endpoint_counts_df %>% arrange(desc(Frequency))

# Lastly I create the bar graph
ggplot(data = endpoint_counts_df, aes(x = reorder(Endpoint, -Frequency), y = Frequency)) + geom_bar(sta
labs(title = "Endpoint Counts",
       x = "Endpoint",
y = "Frequency") + theme_minimal() +
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



Answer: No Observed Effect Level (NOEL) and Lowest Observed Effect Level (LOEL) are the two most common end points used in our database. NOEL is the greatest concentration of a chemical at which no discernible bad effects or substantial changes in the examined organisms are identified over the course of a certain exposure time. It acts as a safety threshold, showing when exposure to a drug does not cause obvious damage or substantial alterations in the organisms under study. LOEL is the lowest dosage or concentration of a drug at which visible adverse effects or substantial changes in the examined organisms are found during a certain exposure time.It

defines the time at which unfavourable effects start to show, indicating that exposure to the drug is hurting or significantly altering the organisms.

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 find the class
class('Litter.data$collectDate')
```

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

```
# It was a character class so i have to convert it to a date
Litter.data$collectDate <- as.Date(Litter.data$collectDate)

class(Litter.data$collectDate)
```

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

```
#Here i use the unique function to find through what dates the litter was sampled as August 2018
unique_dates <- unique(Litter.data$collectDate)
August_2018_dates <- unique_dates[format(unique_dates, "%Y-%m") == "2018-08"]
August_2018_dates
```

```
## [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`?

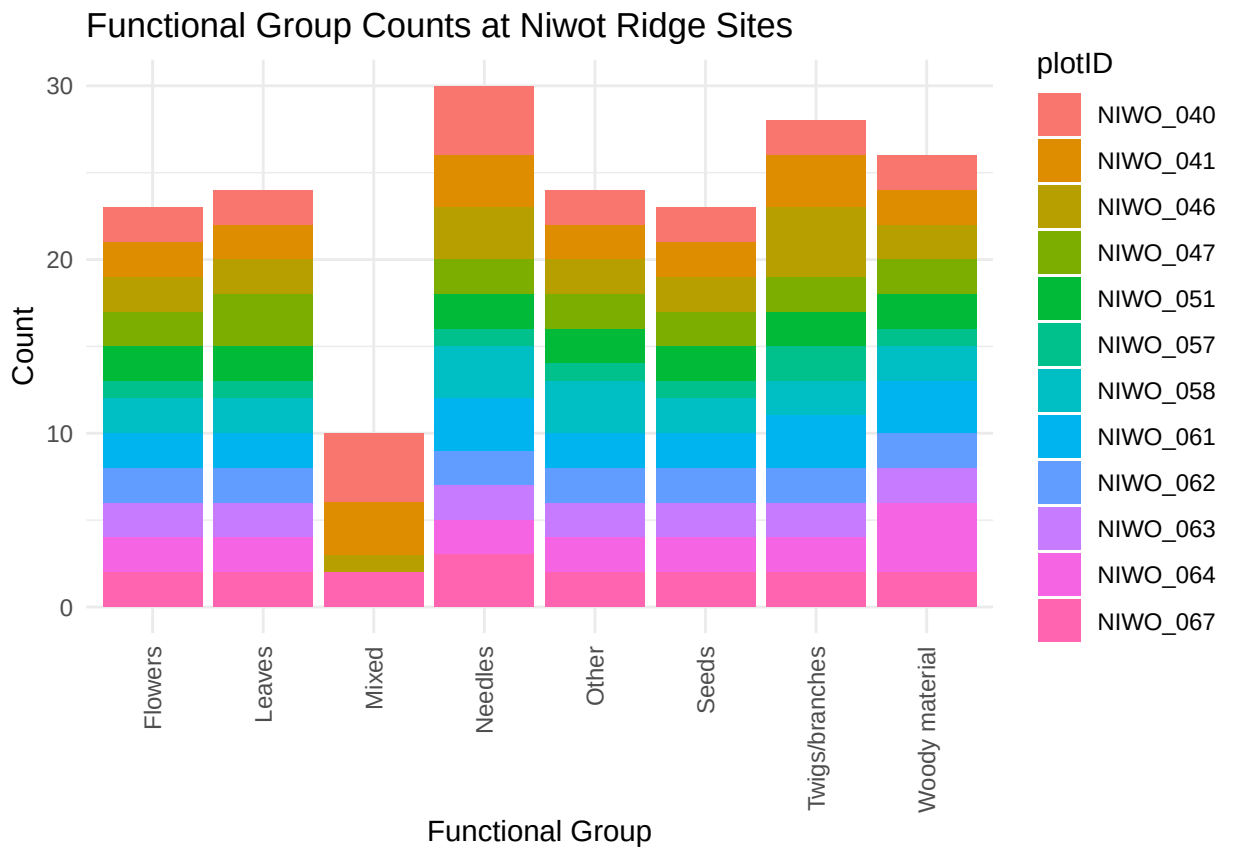
```
unique_plots <- unique(Litter.data$PlotID)
num_plots <- length(unique_plots)
num_plots
```

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

Answer: In order to assist comprehend the distinctive parts existing in my data, unique concentrates on presenting a list of distinct values within a vector or column. On the other hand, summary focuses on producing summary statistics that provide an overview of the distribution and core patterns of numerical data in a data frame.

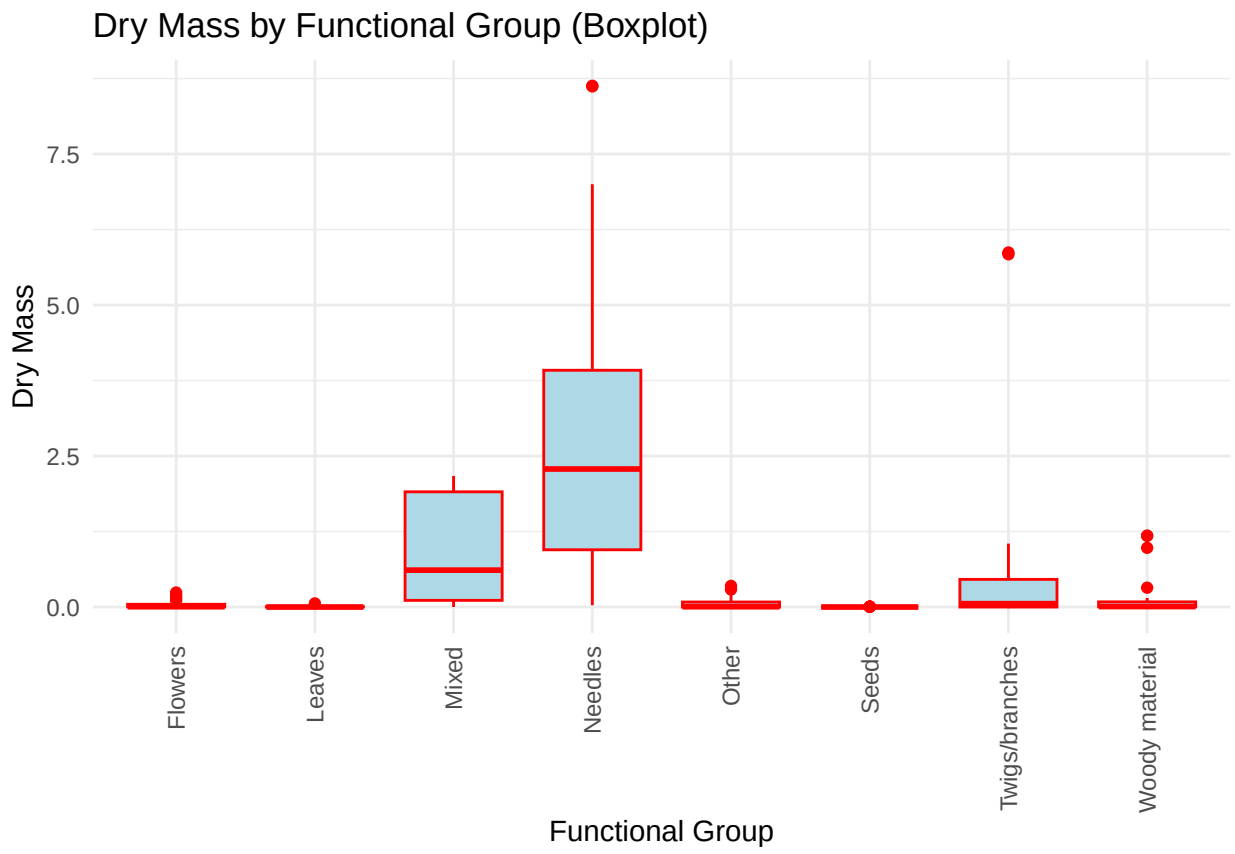
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#Here i use the ggplot command to visually represent the information from the database.
ggplot(Litter.data, aes(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))
```

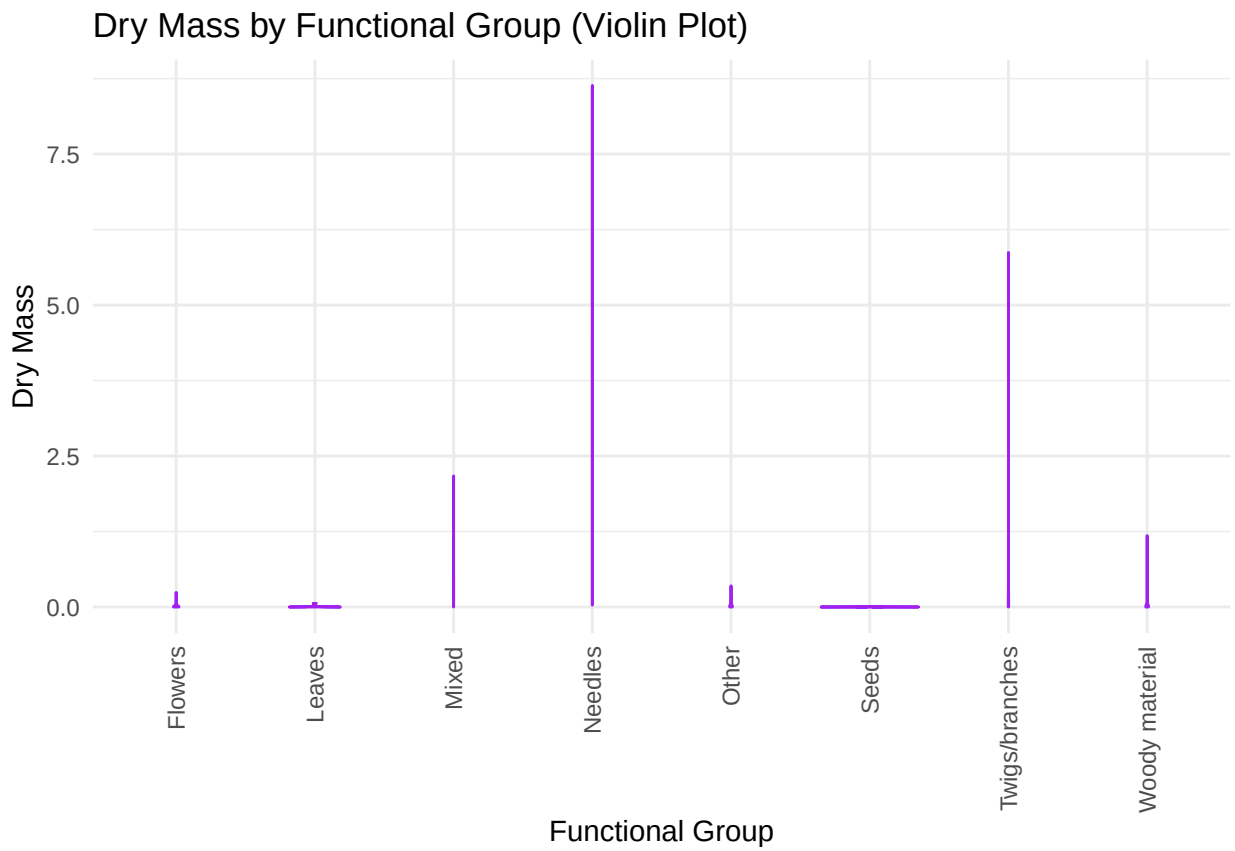


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
# Here i creaate a boxplot for dryMass by functionalGroup
boxplot_plot <- ggplot(Litter.data, aes(x = functionalGroup, y = dryMass)) + geom_boxplot(fill = "lightl
       x = "Functional Group",
       y = "Dry Mass") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
print(boxplot_plot)
```




```
# Now i will create avViolin plot for dryMass by functionalGroup
violin_plot <- ggplot(Litter.data, aes(x = functionalGroup, y = dryMass)) + geom_violin(fill = 'lightgr
       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 investigate the distribution of dry mass values across different functional categories. The boxplot shows the spread (represented by the height of each box representing the interquantile range), the central tendency (represented by the median for each particular functional group as a horizontal line), and the presence of outliers (represented by data points outside the range). Contrarily, the violin plot simply shows the median, which is a distinction we already noticed in the boxplot.

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

Answer: The litters with the highest biomass concentration are Needles and Twing/branches. Disclaimer: I have consulted internet for some of the answers.